European Commission

JRC TECHNICAL REPORTS

Does web anticipate stocks? Analysis for a subset of systemically important banks

Michela Nardo and Erik van der Goot

2014

Report EUR 27023

Joint Research Centre

European Commission
Joint Research Centre
Institute for Protection and Security of the Citizen

Contact information

Michela Nardo
Address: Joint Research Centre, via Fermi 2749 I-21027 Ispra (VA), Italy
E-mail: michela.nardo@jrc.ec.europa.eu
Tel.: +39 0332 78 5968

Erik van der Goot
Address: Joint Research Centre, via Fermi 2749 I-21027 Ispra (VA), Italy
E-mail: erik.van-der-goot@jrc.ec.europa.eu
Tel.: +39 0332 78 5900

JRC Science Hub
https://ec.europa.eu/jrc

JRC 93383
EUR 27023

ISBN 978-92-79-44708-2
ISSN 1831-9424

doi: 10.2788/14130

Luxembourg: Publications Office of the European Union, 2014

Abstract

Is web buzz able to lead stock behavior for a set of systemically important banks? Are stock movements sensitive to the geo-tagging of the web buzz? Between Dec. 2013 and April 2014, we scrape about 4000 world media websites retrieving all public information related to 10 systemically important banks. We process web news with a sentiment analysis algorithm in order to detect article mood. We show that web buzz does not seem to lead stock behavior as Granger test fails to support an average association that goes one-way from web to stocks. We nevertheless find a statistically sound anticipation capacity for single banks with gains ranging from 4 to 12%. Hierarchical clustering and Principal Component Analysis suggest that Euro area level decisions/facts do in fact drive stock behaviour, while web news about single banks only episodically make a difference in stock movements. Our analysis confirms that the location of the web source matters. The use of sources with international echo eliminates some of the noise introduced by irrelevant texts at the country level and improves the predictive power of the model up to 27.5%.

Does web anticipate stocks? Analysis for a subset of systemically important banks

Michela Nardo[1*], Erik van der Goot[2]**

European Commission, Joint Research Centre
Via Fermi 2749 I-21027 Ispra (VA), Italy
([1]) Unit for Financial and Economic Analysis
([2]) Unit for Global Security and Crisis Management

(*) Corresponding author for the paper: michela.nardo@jrc.ec.europa.eu
(**) Corresponding author for EMM: erik.van-der-goot@jrc.ec.europa.eu

Keywords: opinion mining, sentiment analysis, financial news, stock market trends, systemically important banks, pattern recognition, web buzz analysis.
JEL Classification: G10, G14

Abstract

Is web buzz able to lead stock behavior for a set of systemically important banks? Are stock movements sensitive to the geo-tagging of the web buzz? Between Dec. 2013 and April 2014, we scrape about 4000 world media websites retrieving all public information related to 10 systemically important banks. We process web news with a sentiment analysis algorithm in order to detect article mood. We show that web buzz does not seem to lead stock behavior as Granger test fails to support an average association that goes one-way from web to stocks. We nevertheless find a statistically sound anticipation capacity for single banks with gains ranging from 4 to 12%. Hierarchical clustering and Principal Component Analysis suggest that Euro area level decisions/facts do in fact drive stock behaviour, while web news about single banks only episodically make a difference in stock movements. Our analysis confirms that the location of the web source matters. The use of sources with international echo eliminates some of the noise introduced by irrelevant texts at the country level and improves the predictive power of the model up to 27.5%.

Executive summary

The sequence *economic event - investment decisions - stock price jumps* is far from being a stylized fact. Anecdotic comparison of news and movements of the most important stock indices suggests that '*… people sometimes trade on noise as if it were information'* (Black, 1986). Indeed, the majority of the largest stock fluctuations cannot be tied to '*fundamental economic news sufficient to rationalize the size of the observed* [price] *move*' (Cornell, 2013). The crucial question is then how noise influences economic decisions and how noise can be measured.

In the latest years the geometric increase of on line information enabled to address the log-lasting question of stock market predictability from a different perspective, that of Big Data. If the orthodox economic theory postulates that stock markets act erratically and are largely driven by new information unpredictable ex ante, Behavioral Finance suggests instead a certain degree of predictability. Lab experiments and actual observation show that investors are either systematically overconfident in the ability to forecast future stock prices or earnings or subject to waves of optimism and pessimism, causing prices to deviate systematically from their fundamental values. The Big Data perspective has widen the debate, by allowing a broad investigation of human behaviour. On line journals, dedicated blogs, social networks, etc. make possible the access to financial information, even for non-experts, with the advantage of increasing transactions and decreasing costs but with the risk of amplifying rumors, favoring herding behavior and boosting volatility in periods of turbulences. This effect is particularly important for financial systems as perceived weakness of the system could produce a domino effect with dreadful consequences. Yet, the relationship between web buzz and financial movements has to be empirically proven. *Is web buzz indeed able to lead stock behavior?*

We explore this question by relating trading prices and volumes to web buzz for a set of systemically important banks (banks whose performance is crucial for the entire European banking system). Between Dec. 2013 and April 2014 we collected web news using the Europe Media Monitor (EMM), a JRC software that gathers reports from more than 4000 news portals world-wide in 60 languages, classifies the articles, and analyses the news texts (http://emm.newsbrief.eu/overview.html). From text mining we calculate a number of daily web variables, including measures of web "mood", and we associate them to daily stock volumes and trade in different stock markets (both in Europe and in US). The novelty of our analysis is triple. We are the first to analyse the relationship between web buzz and stock behaviour of banks. The possibility to find a statistically significant relationship would help to anticipate turbulences and act accordingly. Secondly, having the complete control of the web data retrieval we can analyse the influence of geo-tagging: *does it make a difference where the news comes from?* The third element of absolute novelty is the multilingual context of our search, as for the same bank tonality is checked in 14 languages. This consents capturing the interaction between the mother company and its branches located in different countries. To the best of our knowledge this has never been done as the large majority of the literature uses English.

Findings

Is web buzz able to lead stock behavior in our dataset? Not on average, according to our data. Granger causality test fails to support an average association that goes one-way from web to stocks. Nevertheless we find a statistically sound anticipation capacity for single banks, particularly Unicredit, Deutsche Bank and Crédit Agricole but also in some cases for BBVA, Royal Bank of Scotland, Société Générale with gains in prediction power ranging from 4 to 12%. The explanation offered by the literature for this poor average performance is that new information is rapidly incorporated into agents' information set so excessive returns rapidly vanish: only very short (ideally intraday) stock price movements can be capitalized. Our analysis confirms the association between web buzz and intraday price movements making this topic a potential candidate for additional research.

Our data indicate that supra-national decisions/facts are driving stock behaviors of banks, while web news about single banks is only episodically making a difference in stock movements. Most likely in these times of financial turbulence announcements of the BCE or of other international authorities are likely to play a crucial role in explaining trade behaviors. In order to capture this effect the construction of a general "sentiment" index will drive our future research efforts.

Are stock movements sensitive to the geo-tagging of the web buzz? We extensively explore different sources of web information distinguishing between local (country) sources and international/world sources. Our data suggest that European stock markets seem to respond to news reported at the international level, rather than locally (i.e., in the country where the bank is located). As EMM is unable to distinguish between "important" and "unimportant" news (nor we know of any text mining algorithm that is able to distinguish information according to the relevance), the use of sources with international echo eliminates some of the noise introduced by irrelevant texts at the country level. Web buzz has a poor association with New York stock data for all banks analysed as overseas stock exchange seems to be guided more by the pattern of the European markets than by that of the web information.

Our analysis does not suggest a clear advantage of measures of web buzz based on tonality (web mood) with respect to other count variables (e.g. relative number of messages). This could be partly due to the algorithm calculating tonality. During the test phase we realized that the tonality failed to identify some important financial news (like for example the downgrade of Deutsche Bank on the 19th of Dec.). Currently the tonality algorithm is being upgraded to provide entity based sentiment. Even so tonality and sentiment analysis on financial texts are the latest and most promising advances in this type of literature.

Contents

List of Figures

List of Tables

Introduction

The geometric increase of on line information enabled to address the log-lasting question of stock market predictability (Barber and Odean, 2001[5]) from a different perspective, that of Big Data. If the Efficient Market Hypothesis (Fama, 1965[20]) postulates that stock markets are largely driven by new information hence their movement is unpredictable ex ante, Behavioral Finance suggests instead a certain degree of predictability (Della Vigna, 2009[17]): investors are found either systematically overconfident in the ability to forecast future stock prices or earnings (Kahneman and Tversky 1979[26]) or subject to waves of optimism and pessimism, causing prices to deviate systematically from their fundamental values (DeBond and Thaler, 1985[15]). On the other hand, sluggish markets, responding only gradually to new information, consent information rents (Chan et al., 1996[9]) and therefore systematic price deviations. The Big Data perspective has reanimated the debate, opening the door to a much broader investigation of human behaviour. According to the standard economic theory extensive web access is likely to favor information spreading and the quick erosion of information (Cornell, 2013[11]; Cutler and Poterba, 1989[12]; Malkiel, 2003[29]). According to the opposite view, web buzz favors herding behavior and boosts volatility in periods of turbulences. On line journals, dedicated blogs, social networks, etc. make possible the access to financial information even for non-experts, amplifying rumors (Shiller, 2000[42]) and facilitating market transactions (Gloor, et al., 2009[23]). This effect is particularly important for financial systems as perceived weakness of the system could produce a domino effect with dreadful consequences. Yet, the relationship between web buzz and financial movements has to be empirically proven.

The literature relating web mining to financial prediction is relatively recent. To the best of our knowledge the first study is due to Wysocki (1998[48]). He proved that, between January and August 1998, the Yahoo! posting volume associated to 50 companies was able to forecast next day trading volumes. About the opposite result, namely internet buzz cannot predict trading volume, is obtained by Tumarkin and Whitelaw (2001[46]) and by Das and Chen (2001[14]), among others. More recently Preis et al. (2012[37]) show that weekly transactions volumes of the companies included in S&P500 are correlated with weekly search volume of the company names. Preis et al. (2013[38]) find that decrease in Dow Jones Industrial Average is preceded by an increase in the search volumes for given financially related terms, and Moat et al. (2013[34]) obtain the same result with Wikipedia views (see Nardo et al., 2014[35], for a survey of stock market predictions using on-line financial news).

Besides financial movements, web mining has been increasingly used as source of information for assessing a wide variety of economic or social phenomena. Web buzz has been proved useful in forecasting box-office revenues (Asur and Huberman, 2010[4]; Doshi et al., 2009[19]; Goel et al., 2010[24]) , movie success (Mishne and Glance, 2006[32]), and videogame sales (Goel et al., 2010[24]). Tweets were considered a reliable alternative to election polls for forecasting the results of 2009 German Federal Election (Tumasjan et al., 2010[47]). Google queries proved to be leading indicators of consumer purchases in selected sectors (automobiles sales, unemployment claims, travel destination planning, and consumer confidence, see McLaren and Shanbhogue, 2011[30]; Choi and Varian, 2012[10]) and blog posts and blog sentiment were related to product sales (Gruhl et al., 2005[25]). More generally, Facebook (Mishne and Rijke, 2006[33]) or Google searches (Preis et al., 2012[37]) have be used to construct macroeconomic indicators correlated with GDP movements.

Here we analyse whether information coming from the web has some predictive power on the stock market behavior for a set of 10 banks considered systemically important by the Financial Stability Board (involving higher loss absorbency requirements): Barclays, BBVA, BNP Paribas, Crédit Agricole, Deutsche Bank, HSBC, Royal Bank of Scotland, Santander, Société Générale and Unicredit. Between Dec. 5th and April 30th 2014 and for each bank we collect daily news coming from more than 4000 electronic media websites worldwide in 60 languages (using Europe Media Monitor, see http://emm.newsbrief.eu). For the same period daily data on stock prices (open, close, highest, lowest) and volumes exchanged are gathered from New York Stock exchange and from various European Stock exchanges (Frankfurt, London, Madrid, Milan, Paris). Two questions guided our analysis: *(i) is web buzz able to lead stock behavior? (ii) Are stock movements sensitive to the geo-tagging of the web buzz?* The relationship between stock data and web news is analysed via cross-correlation function, Granger causality, rank-sum test, Factor and Cluster analysis for each combination of 8 stock prices variables, 12 web buzz variables, 4 set of sources (with different geo-tagging), various stock markets. The novelty of this analysis is triple. We are the first to analyse the relationship between web buzz and stock behaviour of banks. The possibility to find a statistical significant relationship would help to anticipate turbulences and act accordingly. Secondly, having the complete control of the web data retrieval we can analyse the influence of geo-tagging. Google trend does have geo-tagging labels on daily search data, but results are not displayed if the amount of queries is not large enough and even when displayed, the downloadable series are based on random samples of queries (Choi and Varian, 2012[10]; Da et al., 2011[13]). The third element of absolute novelty is the multilingual context of our search, as tonality is checked in 14 languages. This consents capturing the interaction between the mother company and its branches located in different countries. To the best of our knowledge this has never been done as the large majority of the literature uses English (exceptions to the English-oriented analysis are Agić et al. (2010[1]) for Croatian, Remus et al. (2009[39]) and Denecke (2008[18]) for German, and Ahmad et al. (2006[2]) for Chinese and Arabic).

The paper is organized as follows. Section 1 contains a description of the Europe Media Monitor; Section 2 describes the variables used in the analysis. Section 3 presents the results and Section 4 discusses the main research issues and concludes. An Appendix complements the paper describing the methods used and displaying a selection of tables and figures.

1. The Europe Media Monitor

The Europe Media Monitor (EMM) was started in 2002 as a project to support the Commission with its Media Monitoring activities. The main purpose of EMM is to provide monitoring of a large (but selected) set of electronic media, reduce the information flow to manageable proportions by applying categorisation and to provide extra information by analysis of the retrieved texts in the form of entity recognition, entity extraction, recognitions of quotes, sentiment/tonality analysis etc.

EMM is designed as a near real-time monitoring system for new publications. The system generates the required information products continuously and does not rely on (and does not have) a big information archive. Although EMM does maintain an index of all retrieved material, allowing for

limited historical research, the information products always refer to the original publication, mostly on the Internet.

At the core of the EMM system is a processing chain of lightweight extensible processes each running independently and chained together using robust and reliable in-house developed web service architecture. Articles begin their flow through the processing chain as thin RSS (Really Simple Syndication[1]) items that grow as meta-data gets added at each stage of the processing chain.

The first element of this processing chain, the scraper, monitors a number pages/RSS feeds on selected websites for the publication of items and produces a snapshot of all items currently being published on these pages. The selection of websites depends on the information domain to be monitored. For those sites that require 'near real time' monitoring (update frequency measured in minutes), EMM uses a technique which does not rely on 'crawling' the website. Instead, EMM monitors (scrapes) a selected set of HTML pages or RSS feeds on the website. For websites that do not have a clear 'publication' structure the system will crawl the website, but this will reduce the monitoring frequency to a number of times per day.

The second process in the chain receives the snapshot as produced by scraper, and determines the difference between the current snapshot and the previous snapshot (the delta). Based on this delta this process then 'grabs' the new items from the web and extracts the relevant text from the items. For a typical HTML page this is a non-trivial operation as the system tries to identify the 'main article' text from what can be a 'noisy' page. The system then constructs a basic RSS feed, containing the new articles from the source currently being monitored, and adds the extracted text as item metadata. This RSS feed, the basis of the information enhancement and filtering process, is then pushed to the next process in the chain.

Subsequent processes in the chain use the extracted text, and/or metadata added by previous processes, to further enrich the information in the RSS. The Entity Recognition process detects people and organizations in the article from a home grown information base of entities and organizations, populated by an automated (offline) entity recognition system. The next module in the chain performs geo-tagging of the articles, using a multilingual, classified geospatial information base of place names, provinces, regions and countries. The previously recognized entities are used to disambiguate the geo-tags (Clinton is also a place name in Arizona; Paris Hilton is not the Hilton in Paris). Another module extracts quotes from the text and assigns the quotes to the relevant entities in the article. The quote extraction module currently runs in 19 languages.

The tonality/sentiment of an article is determined using 4 sets of 'tonality' words per language, denoting highly positive, positive, negative and highly negative words. These tonality dictionaries are currently available in 14 languages, including the main EU languages (excluding Greek, Hungarian, Bulgarian, Baltic and Scandinavian languages but including Spanish, English, French, German, Dutch, Italian, Check, Slovak and Polish). The total score for an article is calculated by aggregating the score for all tonality words in the article. The score is then transformed using a logarithmic transformation and corrected using a source specific 'tonality bias' which is calculated using a long term 'rolling average' for the source. This ensures that the tonality is as much as possible comparable between sources. The score expresses a full article tonality and is not particularly meaningful as such. For

[1] http://cyber.law.harvard.edu/rss/rss.html

further use, this tonality score is later transferred to any associated categories and aggregated per day. This aggregated value can be used to determine a tonality trend for a category.

The main component that determines the information streams from EMM is a powerful keyword based categorization system. The category definitions allow for word/weight lists, Boolean combinations, proximity and character wildcards. The system deals efficiently with 'overlapping' categories; it is not based on any hierarchical category structure. The system also deals efficiently with languages like Arabic (first character after whitespace is not the first character of the noun) and 'ideograph' languages like Chinese (no whitespace).

The (near) duplicate detection system uses a character trigram signature of the title and description of the articles to calculate a cosine distance measure between an article and all articles in a preceding 24 hour period in the same language. In order to reduce the (potentially huge) number of calculations, the system uses the assigned categories as a way of reducing the set of article signatures used for comparison. The assumption is that (near) duplicate articles share a large set of assigned categories.

Following the duplicate detection system the RSS flows through a second categorization system where new categories are constructed based on the now available article metadata. These new categories are typically defined as the co-occurrence of two or more 'content based' categories and restrictions based on source, language or source country. These new categories are assigned to the articles in an additive way, i.e. the original category information remains. For the purpose of further analysis these new categories are semantically equivalent to the keyword based categories.

All items, now enriched with metadata, are sent on to a number of downstream systems. Some of these downstream systems deal with the individual items, producing RSS feeds per category, per country/category, or sending a mail notifying interested users about new items in a category. All items are indexed to produce a free text searchable index of all articles that entered the system. An analyser module examines the article counts and produces alerts based on deviations from expected daily counts.

The articles also flow into the Clustering and Story Tracking Cache. Every 10 minutes the last 4 hours of articles are hierarchically clustered in every language individually. The clustering process is agglomerative and employs average group linkage to build the clusters using a simple cosine measure to calculate distance. The clustering process continues until the cosine measure falls below a certain set threshold. The article feature vectors are simple word count vectors with some additional ad-hoc rules. Using a sliding window approach the system tracks the evolution of stories over time. This makes it possible to detect 'breaking news', and furthermore to dynamically build (track) very large stories, without having to cluster a huge number of items.

The clustered articles, representing news stories, form the basis of another set of processing modules. These modules are no longer arranged in a pipeline but operate asynchronously and in parallel to each other in order to update the current news story metadata with extra information whenever it becomes available, without delaying the actual 'story'. Examples of these modules are: event metadata extraction, summarization and cross lingual cluster detection.

The results of the information harvesting and processing can be accessed in a number of ways. A website (e.g. http://emm.newsbrief.eu) allows for classical data browsing, and there is a full editorial and publishing system NewsDesk (not publicly accessible) that allows for the creation and publication of high level information products. EMM delivers emails and RSS feeds and there are (free) mobile applications for iPhone, iPad and Android tablets.

Examples of current applications of the EMM technology can be found in different application domains. EMM is used in a number of traditional media monitoring applications by various EU Institutions and Agencies. MedISys (http://medisys.newsbrief.eu) is an instance of EMM specifically developed for internet bio-surveillance and is used by a number of Health Agencies, including the WHO. Open source intelligence for humanitarian and conflict early warning is also covered by at least 3 instances of the EMM system.

At the moment of writing, the publicly accessible instance of EMM, used for the data retrieval described in this paper, monitors around 10000 RSS feeds/HTML pages from 4000 media websites and retrieves and processes around 200.000 new news articles per day. These articles are categorized in around 1500 categories. A selected subset of these categories and the results of the clustering process can be seen on the public EMM website http://emm.newsbrief.eu

2. Variables used in the analysis

From December 5th 2013 to April 30th 2014 we collected all web texts containing the names of the 10 banks considered in the analysis[2] (see Table 1 for an overview of the daily average of texts retrieved). Within the EMM architecture, for each web text we had the possibility to control for the geo-tagging of the source. We classified the web texts in four sets: the complete set of web texts corresponding to all available sources (label ALL), the web texts corresponding to European and USA sources (label EU+US), the web-texts corresponding to European sources only (EU) and the web texts produced by sources located in the country where the bank has its headquarters (labelled Country)[3]. We use geo-tagging to label web articles and analyse whether stock data are more responsive to international news or to local (country) news.

For each bank and each set of sources we compute several summary measures: number of web texts, share of texts with respect to the previous day, share of texts with respect to the total number of articles found by EMM that day, share of texts having positive (negative) tonality, average daily tonality, its standard deviation, polarity, subjectivity and disagreement, where

polarity=(number_pos_ton-number_neg_ton)/(number_pos_ton+number_neg_ton);

[2] Including nicknames, abbreviations and the most common spelling mistakes. We corrected for out of scope texts (e.g., Barclays is the name of a theater and Monte Paschi a basketball team).
[3] The number of sources varies depending on the country: it goes from 134 for Germany, to 48 for Spain. Overall EU + US sources are 1400. We consider both EU and EU+US sources to account for news that could influence the US stock exchange but only indirectly European markets (e.g. in the case of EU banks with branches in the US or in Latin America).

subjectivity=(number_pos_ton+number_neg_ton)/number_articles;

disagreement=(number_pos_ton-number_neg_ton)/number_articles;

Table 1. Daily average of web texts according to the source considered

Daily average number of texts	Sources			
	ALL	EU+US	EU	Country
Deutsche Bank	90	72	61	30
Barclays	71	54	41	11
HSBC	73	46	37	4
Royal B. Scotland	22	18	11	2
BNP Paribas	55	44	38	12
Crédit Agricole	23	19	17	13
Société Générale	42	33	30	16
BBVA	47	28	25	21
Santander	17	11	10	9
Unicredito	41	32	29	14

Polarity expresses whether the daily sentiment is positive or negative while disagreement is a measure of the overall polarity of visions on the daily occurrences. Both should be more related to the positive/negative behavior of stock prices, to price volatility and to the difference between the highest and lowest contracting price. Subjectivity indicates whether a sentiment (no matter its direction) has been expressed and should be more related to the volume exchanged. For missing languages (e.g. Greek, Bulgarian, Swedish, etc.) we set neutral tonality by default. Weekends (and non-contracting days) are excluded to match with stock exchange series.

Daily data on stock prices (open, close, highest, lowest) and volumes exchanged are downloaded from Yahoo! finance for the main contracting markets: New York Stock Exchange (NYSE) for all banks but French and Italian ones (not quoted there) and several European Stock exchanges (Frankfurt for Deutsche Bank, London for HSBC, Royal Bank of Scotland, and Barclays; Madrid for BBVA and Santander; Paris for BNP Paribas, Société Générale, Crédit Agricole and Milan for Unicredit). When web buzz is compared to stock movements in NYSE, the set of web texts is adjusted for the difference in time. Seven summary variables have been constructed from stock data:

(1) close(t)- opening(t); (2) close(t)-close(t-1); (3) w(t)*(close(t)-opening(t)) where w is the volume exchanged in time t divided by the average volume exchanged the previous 5 days; (4) w(t)*close(t)-w(t-1)*close(t-1) ; (5) adjclose(t)-adjclose(t-1) , where adjclose is the close price adjusted for dividends and splits; (6) High(t)-low(t) where High (Low) is the highest (lowest) price reached during the contracting day, this variable is a proxy of the daily price volatility; (6) relative volume exchanged (daily volume divided by the average volume exchanged in the previous 5 days); (7) volume exchanged.

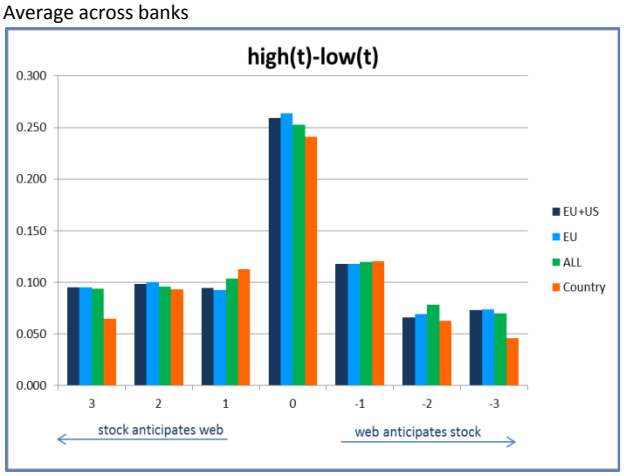
3. Results

3.1 Cross-correlation between web buzz and stock movements

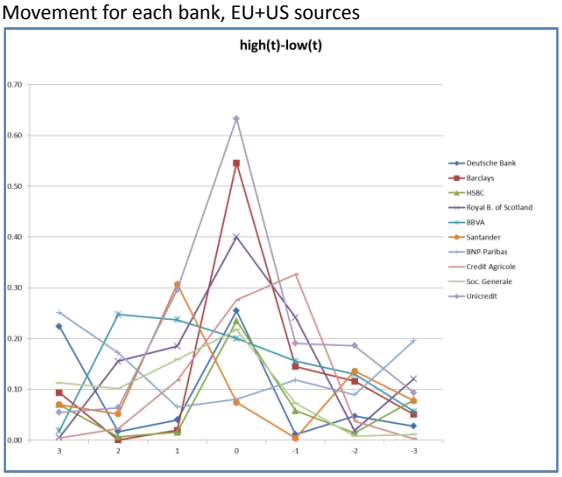
Cross correlation offers a first snapshot of the statistical association between pairs of variables at different points in time (the section on Methods in the Appendix contains the formal definition). In our dataset the largest average cross-correlation between stock variables and web buzz lies between 0.33 and 0.37 at lag δ=0 (contemporaneous correlation), significant at 1% (Table A1 in the Appendix). Checking for individual banks we find correlations up to 0.73 for Barclays and between 0.6 and 0.68 for Unicredit and the Royal Bank of Scotland. Our results are in line with the literature: Gloor et al. (2009[23]) find a positive correlation at δ=0 (highest equals 0.45 significant at 5%) between a set of web variables constructed via semantic social network analysis and the prices of 21 stocks. Significant cross correlation (around 0.3) is found between search data and volume traded for some specific terms and only for instantaneous correlation by Preis et al. (2010[36]), while Bordino et al. (2012[8]) find on average 0.31 at δ=0 with peaks of 0.83 when calculating cross-correlation between trading volumes and Yahoo! queries for a sample of 87 companies in NASDAQ100.

Figure 1. Cross correlation function between Number of web texts and price volatility (highest minus lowest daily price). European stock exchange data.

Average across banks



Movement for each bank, EU+US sources



In our dataset the highest cross-correlation is at δ=0 and usually corresponds to trading volumes rather than trading prices (Table A1 in the Appendix), again in line with the literature (Bordino et al., 2012[8]; Preis et al., 2010[36]; Ruiz et al., 2012[40]). Among the web variables considered, the *(relative) number of articles* and the *number of articles with a given tonality* are found as those displaying the highest correlation with stock prices and volumes. Other measures based on tonality (polarity or disagreement) display much lower correlation. This result holds for all subset of sources considered.

Beyond trading volumes, the difference between the highest and the lowest trading price (a proxy of daily trading volatility) seems to be related to web buzz, with an average correlation across banks of 0.26 and peaks of 0.63 (Figure 1). This observation is, again, in line with the literature suggesting a positive relationship between web news and volatility. Gidófalvi (2001[21]) finds significant

correlation between stock prices (for a set of 12 companies) and news articles 20 minutes before/after the news is made public. Increased trade (and increased returns) as the synchronous trading increases is found by Saavedra et al. (2011[41]).

3.2 Are stock movements sensitive to the geo-tagging of the web buzz?

Table 2. Granger causality test between opening prices in NY stock exchange (nyse) and national stock exchanges (London ,Madrid, Frankfurt). Various banks.

Pairwise Granger Causality Tests Lags: 3 Null Hypothesis	F-Statistic	Probability
BARCLAYS_NYSE does not Granger Cause BARCLAYS_LSE	0.821	0.486
BARCLAYS_LSE does not Granger Cause BARCLAYS_NYSE	38.702	0.000
BBVA_NYSE does not Granger Cause BBVA_MAD	0.365	0.778
BBVA_MAD does not Granger Cause BBVA_NYSE	25.472	0.000
DEUTSCHE_BANK_NYSE does not Granger Cause DEUTSCHE_BANK_FRAN	1.791	0.157
DEUTSCHE_BANK_FRAN does not Granger Cause DEUTSCHE_BANK_NYSE	26.431	0.000
HSBC_NYSE does not Granger Cause HSBC_LSE	2.207	0.094
HSBC_LSE does not Granger Cause HSBC_NYSE	22.459	0.000
RBS_NYSE does not Granger Cause RBS_LSE	0.905	0.443
RBS_LSE does not Granger Cause RBS_NYSE	34.964	0.000
SANTANDER_NYSE does not Granger Cause SANTANDER_MAD	0.667	0.575
SANTANDER_MAD does not Granger Cause SANTANDER_NYSE	22.631	0.000

Note: For each bank in the sample and each web-variables we estimate two equations. (1) web anticipates stocks: $S_t = \alpha + \beta_i S_{t-i} + \beta_2 W_t + \beta_{2+i} W_{t-i} + \varepsilon_t$ and (2) Stocks anticipate web: $W_t = \alpha + \beta_i W_{t-i} + \beta_2 S_t + \beta_{2+i} S_{t-i} + \varepsilon_t$ for *i=1,…,6*. We use as stock variables trade prices, volumes and volatility. Non-significant variables are discarded using Bayesian information criterion. Unit root is tested beforehand with the Augmented Dickey-Fuller test and first-order series are calculated when unit root is not rejected. Residuals have been checked for normality with Jarque-Brera test and for serial correlation and more general ARCH effects with the Breusch-Godfrey test and with the ARCH LM test respectively. In all reported cases, unit root, heteroschedasticity, non-normality and ARCH effects are discarded.

Web buzz seems to have a poor association with New York stock data for all banks analysed: no matter which set of web sources is considered, cross correlation is systematically lower when New York stock data are used (Appendix, Table A1, left hand side). We further explore the issue regressing NYSE returns (and volumes) onto its past values and on present and past values of web buzz. The web variables almost always result to be non-significant. A further look to the data

confirms that New York stock values reacts much more to the corresponding movements in European stocks (NYSE opens 5/6 hours later) than to web buzz, no matter where this buzz comes from. The correlation between opening prices ranges from 0.91 to 0.98 for all banks considered. A Granger causality test on opening prices clearly confirms that association goes one-way from European to NY stock exchanges (Table 2).

Our analysis shows that the location of the source matters (Table 3). Web buzz derived from EU+US sources or from world sources improves the predictive power of a regression of stocks onto web buzz up to 27.5%, if compared to the same regression but with web buzz obtained from Country sources. Modest gains from considering a wide range of international sources are obtained for Crédit Agricole and Santander. The only outlier is HSCB where the web buzz calculated from EU+US sources seems to have (26.5%) less predictive power than that obtained from UK sources. However, the low number of daily UK texts extracted (Table 1) limits the relevance of this result. For this bank, the use of all web sources, while doubling the number of daily texts processed, improves the R^2 only by 1.56% as compared to the use of EU sources only. The explanation of this poor performance comes from the amount of irrelevant information (false positive) still present in HSBC data and influencing the quality of results.

Google trend has geo-tagging labels on daily search data, but results are not displayed if the amount of queries is not large enough and even when displayed, the downloadable series are based on random samples of queries.

Table 3. Percentage difference in R-square according to the source of the web buzz.

	sources		
	EU-US vs Country	All vs Country	EU vs Country
Barclays	24.1%	23.0%	21.8%
BBVA	21.2%	5.2%	17.1%
BNP Paribas	24.6%	29.8%	23.9%
Crédit Agricole	3.3%	1.3%	2.5%
Deutsche Bank	22.4%	24.7%	22.0%
HSBC	-26.5%	-26.9%	-28.9%
Royal B. Scotland	27.5%	29.7%	29.5%
Santander	4.8%	-0.2%	3.5%
Société Générale	11.1%	2.4%	6.1%
Unicredito	14.1%	14.9%	11.6%

Note: We estimate the equation: $S_t = \alpha + \beta_1 S_{t-1} + \beta_2 W_t + \beta_3 W_{t-1} + \varepsilon_t$ for each bank in the sample and each of the four different information sets for the web buzz (*W* denotes web variables and *S* stock variables). Web variables are calculated from web texts coming from: 1) a source located in the USA and in the European Union (EU+US); 2) a source located in the European Union (EU); 3) sources all over the world (ALL); 4) sources located in the country where the bank has its headquarters (Country). For each estimated model we calculate the percentage change in the model fit (R^2) using option 4 as baseline.

Overall European stock markets seem to respond to news reported at the international level, rather than locally. The importance of the news is probably the explanation. Main news, those more likely to drive stock prices, is also those actually reported by the international (financial) journals. As EMM

is unable to distinguish between "important" and "unimportant" news (as soon as the required keywords are in it) the information reported at the international level is a synonymous of financial relevance hence more likely to be related to stock prices. Indeed, repeating the analysis only on the set of articles with international echo (i.e. on the difference between the set of articles labeled *EU+US* or *ALL* and those labeled *Country*) we obtain similar results for the cross correlation (Table A5 in the Appendix): the average cross correlation looks pretty much the same with or without Country sources. The Granger test (Table B5 in the Appendix) confirms the similarities. Overall the use of web texts with international echo seems to eliminate some of the noise introduced by irrelevant texts at the country level.

3.3 Does web buzz lead stock behavior?

While cross-correlation only supplies a fist idea of the relationship between two variables, Granger causality test helps verifying the bivariate dependency structure of the data (see the section on methods for a discussion). We run the Granger test for each pair of web buzz and stock variable and each set of geo-tagged sources. The hypothesis that web buzz on average anticipates stock movements, receives little support from our data on continental stock exchanges. Table B1 in the Appendix shows that no matter the geo-tagging of the sources, the direction (Web vs. Stock) never obtains much stronger support as compared to the opposite direction (Stock vs. Web) for returns, volatility and volumes, at least on average across all 10 banks considered. The literature is supporting our findings. Gilbert and Karahalios (2010[22]) and Bollen et al. (2011[7]) find that web information is most likely causing price movements than the reverse. No prediction power for stock prices or volatility is found by Antweiler and Frank (2004[3]) with Naïve Bayesian machine learning. De Choudhury et al. (2008[16]) with Support Vector Machine, find out that only after the occurrence of "big" events web mining shows explanatory power (up to 87%).

When looking at individual banks (Appendix, Table B2) web buzz does not anticipate stock movements for Barclays, HSCB, Santander and BNP Paribas. For BBVA, Société Générale, and Royal Bank of Scotland results are more positives while a certain degree of anticipation is found for web news related to Crédit Agricole, Unicredit, and Deutsche Bank. For those banks the direction (W vs. S) obtains more support (significant at 1%) than the opposite direction (S vs. W). A closer look to each bank shows that there is no general pattern in the predicted stock variable: while for Deutsche Bank and Unicredit web buzz anticipates both stock returns and volumes (significant at 1% level), for Crédit Agricole web buzz mainly anticipates volatility (significant at 1%) while for BBVA, Société Générale, and Royal Bank of Scotland the only anticipatory power (significant only at 5% level) is on stock returns (Table 4).

When web buzz clearly leads stock movements the gain in model fit (average reduction of the Residual Sum of the Squares) is between 6% and 7.8% for Deutsche Bank, up to 9% - 12% for Crédit Agricole and Unicredit respectively. These figures are higher than the 5% of Bordino at al. (2012[8]), Lavrenko et al.(2000[27], [28]) and Mittermayer (2004[31]) who finds that an intraday trading strategy that uses web buzz can produce an average gain ranging between 0.1% to 0.5% with respect to a random strategy with zero expected gain.

Table 4. Granger Causality test.

Result (p-values) for a subsample of banks. W vs. S corresponds to the null hypothesis H0: web does not Granger cause stocks, while S vs. W corresponds to the H0: stock does not Granger cause web.

bank	web variable	stock variable	p-value of W vs. S	p-value of S vs. W
Crédit Agricole	average_tonality	close(t)-close(t)	0.0403	0.2760
		close(t)-close(t-1)	0.0231	0.0699
	number of web texts	high(t)-low(t)	0.0032	0.1890
Deutsche Bank	number of web texts	w(close(t)-close(t-1))	0.0095	0.3174
		volume exchanged	0.0114	0.2267
	number_neg_ton	w(close(t)-close(t-1))	0.0103	0.3178
		volume exchanged	0.0159	0.2252
	number_pos_ton	w(close(t)-close(t-1))	0.0059	0.4399
		volume exchanged	0.0062	0.3540
Unicredit	number of web texts	w(close(t)-close(t-1))*	0.0000	0.3478
		volume exchanged	0.0000	0.2588
	number_neg_ton	w(close(t)-close(t-1))*	0.0001	0.3737
		volume exchanged	0.0009	0.2223
	number_pos_ton	w(close(t)-close(t-1))*	0.0020	0.4485
		volume exchanged	0.0044	0.4362
BBVA	number_neg_ton	w(close(t)-close(t-1))	0.0094	0.0791
	number_pos_ton	close(t)-open(t)	0.0288	0.1443
		close(t)-close(t-1)	0.0369	0.2965
Royal Bank of Scotland	number of web texts	w(close(t)-close(t-1))	0.0000	0.4561
Société Générale	number_pos_ton	w(close(t)-close(t-1))	0.0223	0.5670
	web(t)/web(t-1)	close(t)-close(t-1)	0.0093	0.1155

(*) Unicredit: for all web variables corresponding to this stock variable the Granger test points to " web anticipates stock". Here only the results with the highest F are reported.

Note: For each bank in the sample and each web-variables we estimate two equations. (1) web anticipates stocks: $S_t = \alpha + \beta_i S_{t-i} + \beta_2 W_t + \beta_{2+i} W_{t-i} + \varepsilon_t$ and (2) Stocks anticipate web: $W_t = \alpha + \beta_i W_{t-i} + \beta_2 S_t + \beta_{2+i} S_{t-i} + \varepsilon_t$ for *i=1,…,6*. We use as stock variables trade prices, volumes and volatility. Non-significant variables are discarded using Bayesian information criterion. Unit root is tested beforehand with the Augmented Dickey-Fuller test and first-order series are calculated when unit root is not rejected. Residuals have been checked for normality with Jarque-Brera test and for serial correlation and more general ARCH effects with the Breusch-Godfrey test and with the ARCH LM test respectively. In all reported cases unit root, heteroschedasticity, non-normality and ARCh effects are discarded.

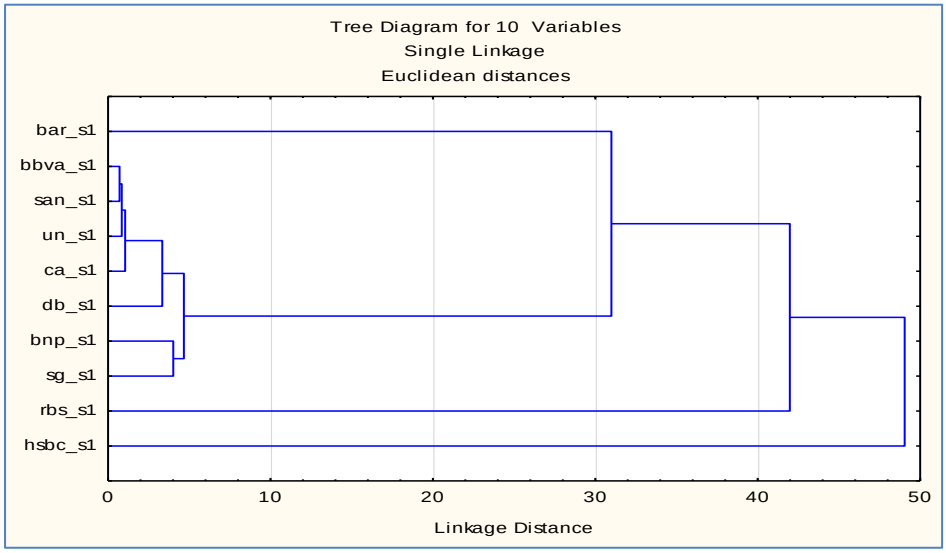
The average results of Granger causality test are weaker when NYSE data are considered (Appendix, Tables B3 and B4). The relationship between web buzz and stock movements is virtually non-existing, reinforcing the idea that news about European banks, if at all, they shake European stock exchanges. When some anticipatory effect is present in NYSE it is essentially for Santander, Barclays, and Royal Bank of Scotland, and mostly on stock prices (significant at 1% for the first two banks) and on volatility (at 5% for the RBS). The gains in terms of model fit are lower, ranging between 4 and 5%.

We are fully aware that the Granger causality test is valid under the assumption of normality of error terms and the linearity of the model. Normality, tested with the Jarque-Brera test, is rejected only in few cases, usually for the stock variable high(t)-low(t). To account for non-normality and for non-linearities we perform the U-Rank test. Table B6 in the Appendix reports the results of the U-test for 10, 50, 100 and 1000 bootstraps on the estimated residuals. We notice the general tendency of this test to refuse the rejection of the null as the number of bootstraps increases. Overall, U-rank test confirms the finding obtained with Granger: the failure of web buzz to lead, on average, stock movements. Nevertheless we find cases in which a clear leading role of web information is positively assessed (for those cases web anticipate mostly trade prices). In particular for BBVA and Deutsche Bank and for the cases in which Granger residuals did not fulfill the normality assumptions, U-test do not reject the leading role of web information on volatility.

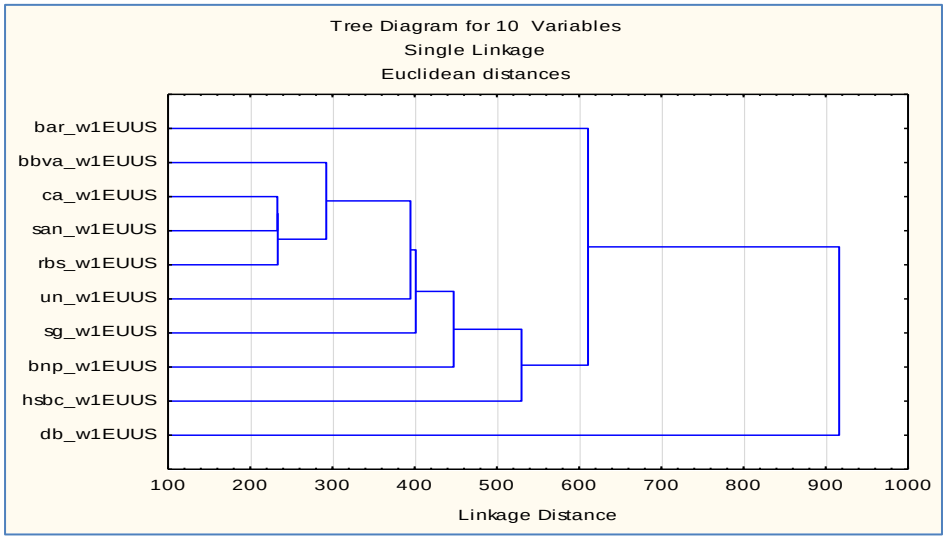
In our dataset we observe high correlations between stock volumes and prices of some banks, e.g. BBVA has a contemporaneous correlation of 0.74 (for volumes) and 0.9 (for prices) with Santander, and Société Générale has a price correlation of 0.84 and 0.8 with BNP and Crédit Agricole respectively. This induced us to explore the hypothesis of a geographical political/economic clustering not captured by the web buzz variables used here. A simple hierarchical clustering on the price variables Close(t)-Open(t) (Figure 2.1) shows, in fact, that euro-area banks tend to cluster together very fast while English banks are far apart and move differently (the same happens for the variable $high_t$-low_t, our proxy of volatility). However, this is not the case for web buzz variables (Figure 2.2), where the differentiation between continental and UK banks is not clearly defined (Royal Bank of Scotland clusters quickly with Santander and Crédit Agricole while Deutsche bank behaves differently and is clustered at the very end).

Figure 2. Hierarchical Cluster analysis, all banks*

2.1 Stock price variable s1: (*Close(t)-Open (t)*)



2.2 web buzz variable w1: *Number of articles* (EU+US sources)



(*) Banks: Barclays (bar), BBVA (bbva), Santander (san), Unicredit (un), Crédit Agricole (ca), Deutsche Bank (db). BNP Paribas (bnp), Société Générale (sg), Royal Bank of Scotland (rbs), HSBC (hsbc).

To explore further the issue we use Principal Component Analysis (PCA) on stock prices and volatility. We find that while euro area banks are all robustly loaded (with the same sign) by the same single factor (with the caveat of Unicredit on the variable *high(t) –low(t)*), UK banks tend to be loaded by multiple factors (especially HSCB which stands out as the most *diverse* bank, Table C1 in the Appendix). Euro area banks show a unique common driver explaining 74.06% of the total euro area variance, all the remaining variance practically represents idiosyncratic bank-related noise (Appendix, Table C2). If web buzz were to reflect/anticipate stock movements we should expect a grouping in the PCA similar to than found for the stock variables. This is not the case: the PCA on the web variable *Number of articles* reveals at least 5 different (orthogonal) relevant factors, the first of which explaining only 15.55% of the total variance (the first PCA factor on the stock variable represents about 60% of the total variance). A possible explanation is that news for one bank pushes bear/bull reactions on related banks (we notice co-movements at the country level). In order to verify this hypothesis intra-daily stock data would be needed. Another possible explanation is that Euro area decisions/facts are in fact driving stock behaviors, while web news about single banks is only episodically making the difference in stock movements.

We further investigated if different drivers of web information could explain the common driver of stocks by regressing the first factor of the PCA done on stocks onto the first 3 factors obtained from the PCA on the web-variable (three is the number of eigenvalues higher than 1). Results are disappointing (Table 5) as all estimated parameters for web variables are not significantly different from zero. Results therefore seem to confirm that web information is not, on average, able to capture the main trend associated to stocks. Results do not say that web buzz could not be relevant in explaining stock behavior but rather than web buzz about individual banks cannot. Yet, we believe that the web hosts valuable information thus in future works we will investigate general economic/financial trends based on web information.

Table 5. Is web buzz able to explain common drivers in stocks? Regression Results of PCA factors.

PCA_s_f1 is the first factor of the PCA on the variable *close(t)-open(t)*, euro-area banks, country stock exchange data. Variables PCA_w_F1 to 3 are the first 3 factors of the PCA on the web variable *number of web texts* (EU+US sources), euro area banks.

Dependent Variable: PCA_S_F1				
Sample: 1 97				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.95E-17	0.234	-2E-16	1
PCA_W_F1	0.0173	0.193	0.090	0.929
PCA_W_F2	-0.1264	0.208	-0.607	0.546
PCA_W_F3	0.0323	0.221	0.147	0.884
R-squared	0.004256	Mean dependent var		-8.24E-17
Adjusted R-squared	-0.027864	S.D. dependent var		2.276854
S.E. of regression	2.308357	Akaike info criterion		4.551312
Sum squared resid	495.5517	Schwarz criterion		4.657486
Log likelihood	-216.7386	F-statistic		0.132511
Durbin-Watson stat	1.943029	Prob(F-statistic)		0.940488

4. Discussion and conclusions

The sequence *economic event - investment decisions - stock price jumps* is far from being a stylized fact. Anecdotic comparison between economic news and ex-post movement in aggregated stock prices claims that the majority of the largest movements in S&P500 (Cutler et al., 1989[12]) and CRSP Total Market Index cannot be tied to *fundamental economic news sufficient to rationalize the size of the observed* [price] *move* (Cornell, 2013[11]). As stated by Black (1986[6]): *"… people sometimes trade on noise as if it were information"*. Is web buzz an ingredient of the missing link of this sequence?

We explore this hypothesis by relating trading prices and volumes to web buzz for a set of systemically important banks, namely Barclays, BBVA, BNP Paribas, Crédit Agricole, Deutsche Bank, HSBC, Royal Bank of Scotland, Santander, Société Générale and Unicredit. Web buzz is obtained by monitoring more than 4000 media websites from all over the world and extracting the texts containing an exogenously supplied set of bank related keywords. From text mining we calculate a number of web variables and we associate them to stock volumes and trade in different stock markets. We compute cross-correlation, Granger causality and U-rank tests. We use Cluster and Principal Component Analysis to investigate the structure of the data set.

Is web buzz able to lead stock behavior in our dataset? Not on average, according to our data. Granger test fails to support an average association that goes one-way from web to stocks. Nevertheless we find a statistically sound anticipation capacity for single banks, particularly

Unicredit, Deutsche Bank and Crédit Agricole but also in some cases for BBVA, Royal Bank of Scotland, Société Générale with gains in RSS ranging from 4 to 12%. The explanation offered by the literature for this poor average performance is that new information is rapidly incorporated into agents' information set so excessive returns rapidly vanish: only very short (ideally intraday) stock price movements can be capitalized (Schumaker and Chen, 2006[43], 2009[44])'. In our analysis, cross correlation and in some cases U-rank test confirm the association between web buzz and intraday price movements making this topic a potential candidate for future research.

Our data indicate that supra-national decisions/facts could drive stock behaviors, while web news about single banks is only episodically making a difference in stock movements. Most likely in these times of financial turbulence announcements of the BCE or of other international authorities are likely to play a crucial role in explaining trade behaviors. In order to capture this effect the construction of a general "sentiment" index will drive our future research efforts.

Are stock movements sensitive to the geo-tagging of the web buzz? We extensively explore different sources of web information distinguishing between local (country) sources and international/world sources. Our data suggest that European stock markets seem to respond to news reported at the international level, rather than locally (in the country where the bank is located). As EMM is unable to distinguish between "important" and "unimportant" news (nor we know of any text mining algorithm that is able to distinguish information according to the relevance), the use of sources with international echo eliminates some of the noise introduced by irrelevant texts at the country level.

Web buzz has a poor association with New York stock data for all banks analysed as overseas stock exchange seems to be guided more by the pattern of the European markets than by that of the web information.

Our analysis does not suggest a clear advantage of measures of web buzz based on tonality with respect to other count variables (e.g. relative number of messages). This could be partly due to the algorithm calculating tonality. During the test phase we realized that the tonality failed to identify some important financial news (like for example the downgrade of Deutsche Bank on the 19th of Dec.). Currently the tonality algorithm is being upgraded. Even so tonality and sentiment analysis on financial texts are the latest and most promising advances in this type of literature (see Das and Chen, 2001[14]; Tulankar et al., 2013[45]; Zhang et al., 2010[49]; Zhai et al., 2011[50]). Finally another limitation of our analysis is surely the restricted set of bank analysed. Enlarging the group of banks would lead us to face the tradeoff between wide coverage but lower number of daily web texts extraction (e.g. we obtain very few texts and not every day for e.g. the Finnish Pohjola and the Belgian KBC). Aggregation at the weekly level could be a solution worth exploring.

References

1. Agić, Z., Ljubešić, N., Tadić, M. Towards Sentiment Analysis of financial texts in Croatian. LREC *European Language Resource Association,* 2010.
2. Ahmad, K., Cheng, D., Almas, Y. Multi-Lingual Sentiment Analysis of Financial News Stream. *Proceeding of the conference Grid Technology for Financial Modeling and Simulation*, Palermo, Italy, February 3-4, 2006.
3. Antweiler, W. & Frank, M.Z. Is All that Talk Just Noise? The Information Content of Internet Stock Message Boards. *Journal of Finance* **59(3)**, 1259-1294 (2004).
4. Asur, S. & Huberman, B. A. Predicting the Future with Social Media. (2010) http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf (Accessed: 16th October 2014)
5. Barber, B. & Odean, T. The Internet and the Investor. *Journal of Economic Perspectives*, **15(1)**, 41-54 (2001).
6. Black, F. Noise. *Journal of Finance* **41(3)**, 528-543 (1986).
7. Bollen, J., Mao, H. & Zeng, X. Twitter Mood Predicts the Stock Market. *Journal of Computational Science* **2**, 1-8 (2011).
8. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. & Weber, I. Web Search Queries Can Predict Stock Market Volumes? *PLloS One* **7(7)**, e40014 (2012).
9. Chan, L. N., Jegadeesh, N. & Lakonishok, J. Momentum Strategies. *Journal of Finance* **51(5)**, 1681-1713 (1996).
10. Choi, H. & Varian, H. Predicting the Present with Google Trends. *Economic Record* **88**, 2-9 (2012).
11. Cornell, B. What Moves Stock Prices? Another Look. *Journal of Portfolio Management* **39(3)**, 32-38 (2013).
12. Cutler, D., Poterba, J., & Summers, L. What Moves Stock Prices? *Journal of Portfolio Management* **15**, 4-12 (1989).
13. Da, Z., Engelberg, J., Gao, P. In Search of Attention. *Journal of Finance* **665**, 1461-1499 (2011).
14. Das, S.R. & Chen, M. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *EFA 2001 Barcelona Meetings,* 2001. Available at SSRN: http://ssrn.com/abstract=276189 or http://dx.doi.org/10.2139/ssrn.276189 (Accessed: 16th October 2014).
15. DeBond, W. F. M. & Thaler, R. Does the Stock Market Overreact? *Journal of Finance* **40**, 793-805 (1985).
16. De Choudhury, M., Sundaram, H., John, A., & Seligmann, D.D. Can Blog Communication Dynamics be Correlated with Stock Market Activity? In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedial*, Pittsburgh, 2008.
17. Della Vigna, S. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature* **47(2)**, 315–372 (2009).
18. Denecke, K. Using SentiWordNet for Multilingual Sentiment Analysis. *Data Engineering Workshop*, IEEE 24th International Conference, 2008.
19. Doshi, L., Krauss, J., Nann, S. & Gloor, P. Predicting Movie Prices through Dynamic Social Network Analysis. *Proceeding Collaborative Innovations Network Conference* (COINs 2009), Savannah, 2009.
20. Fama, E. The Behavior of Stock Market Prices. *The Journal of Business* **38(1)**, 34-105 (1965).

21. Gidófalvi, G. Using News Articles to Predict Stock Market Movements. *Mimeo*, University of California San Diego: Department of Computer Science and Engineering, 2001.
22. Gilbert, E. & Karahalios, K. Widespread Worry and the Stock Market. *Fourth International AAAI Conference on Weblogs and Social Media*, *ICWSM,* 2010.
23. Gloor, P., Krauss, J., Nann, S., Fischbach, K. & Schroder, D. Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. *International Conference on Computational Science and Engineering*, Vancouver, 2009.
24. Goel, S., Hofman, J., Lahaie, S., Pennock, D. & Watts, D. Predicting Consumer Behavior with Web Search. *PNAS* **107(41)**, 17486–17490 (2010).
25. Gruhl, D., Guha, R., Kumar, R., Novak, J. & Tomkins, A. The Predictive Power of Online Chatter. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international Conference on Knowledge Discovery in Data Mining*, New York, 78–87 (2005).
26. Kahneman, D. & Tversky, A. Prospect Theory: an Analysis of Decision Under Risk. *Econometrica* **47(2)**, 263-291 (1979).
27. Lavrenko, V. M., Schmill, M., Lawrie, D. & Ogilvie, P. Mining of Concurrent Text and Time Series. *6th ACM SIGKDD International Knowledge Discovery and Data Mining*, Boston, 2000.
28. Lavrenko, V. M., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J. Language Models for Financial News Recommendations. *Proceedings of the 9th International Conference on Information and Knowledge Management*, Washington DC, 2000.
29. Malkiel B.G. The Efficient Market Hypothesis and its Critics. *Journal of Economic Perspectives* **17(1)**, 59-82 (2003).
30. McLaren, N. & Shanbhogue, R. Using Internet Search Data as Economic Indicators. *Bank of England Quarterly Bulletin* **51(2)**, 134-140 (2011).
31. Mittermayer, M. A. Forecasting Intraday Stock Price Trends with Text Mining Techniques. *Proceeding of the 37th Hawaii International Conference on System Sciences*, Hawaii, 2004.
32. Mishne, G. & Glance, N. Predicting Movie Sales from Blogger Sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, Stanford University, 2006.
33. Mishne, G. & Rijke, M. Capturing Global Mood Levels Using Blog Posts. In AAAI *2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, Stanford University, 2006.
34. Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E. & Preis, T. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports* **3** Article number: 1801 (2013).
35. Nardo, M., Petracco-Giudici, M. & Naltsidis, M. Walking down Wall Street with a tablet. A survey of stock market predictions using the web. *Mimeo, JRC* (2014).
36. Preis, T., Reith, D. & Stanley, H. E. Complex Dynamics of our Economic Life on Different Scales: Insights from Search Engine Query Data. *Philosophical Transactions of the Royal Society A* **368**, 5707-5719 (2010).
37. Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. Quantifying the Advantage of Looking Forward. *Scientific Reports* **2**, Article number: 350 (2012).
38. Preis, T., Moat, H. S. & Stanley, H. E. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports* **3**, Article number: 1684 (2013).
39. Remus, R., Heyer, G., Ahmad, K. Sentiment in German Language News and Blogs, and the DAX. *Text Mining Services 2009*, Leipziger Beitrage zur Informatik, pp. 149-158, Leipzig, Germany, 2009.

40. Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. Correlating Financial Time Series with Micro-Blogging Activity. In Adar, E., Teevan, J., Agichtein, E. & Maarek, Y. (Eds.) *Proceedings of the Fifth International Conference on Web Search and Web Data Mining*, 513-522 (2012).
41. Saavedra, S., Hagerty, K. & Uzzi, B. Synchronicity, Instant Messaging, and Performance among Financial Traders. *PNAS Early Edition*, Edited by S. L., Levin, Princeton University, Princeton NJ, 2011.
42. Shiller, R. J. *Irrational Exuberance*. Princeton: Princeton University Press, Princeton NJ, 2000.
43. Schumaker, R. & Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: the AZFinText System. *12th Americas Conference on Information Systems*, 2006.
44. Schumaker, R. & Chen, H. A Quantitative Stock Prediction System Based on Financial News. *Information Processing and Management* **45(5)**, 571-583 (2009).
45. Tulankar, S., Athale, R. & Bhujbal, S. Sentiment Analysis of Equities using Data Techniques and Visualizing the Trends. *Int. Journal of Computer Science Issues* **10(4)**, 265-269 (2013).
46. Tumarkin, R. & Whitelaw, R. F. News or Noise? Internet Message Board Activity and Stock Prices. *Financial Analysts Journal* **57**, 41-51 (2001).
47. Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. Predicting Elections with Twitter: what 140 Characters Reveal about Political Sentiment. *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.
48. Wysocki, P. D. Cheap Talk on the Web: The Determinants of Posting on Stock Message Boards. *Working Paper n. 98025*, University of Michigan Business School (1998).
49. Zhang, X., Fuehres, H. & Gloor, P. Predicting Stock Market Indicators through Twitter ' I hope it is not as bad as I fear'. *Procedia – Social and Behavioral Science*, 2010.
50. Zhai, J., Cohen, N. & Atreya, A. Sentiment Analysis of News Articles for Financial Signal Prediction. *Mimeo*, University of Stanford, 2011.

Acknowledgements: we tank S. Lechner and M. Petracco-Giudici for comments and suggestions; M. Naltsidis and all EMM team for technical support.

Author Contributions

E.v.d.G. and his team implemented EMM search and text mining and wrote the section on EMM functioning in the Supplementary Information. M.N. performed all the calculations and wrote the remaining sections.

Disclaimer

The views expressed in this paper are purely those of the writers and may not in any circumstances be regarded as stating an official position of the European Commission.

APPENDIX

A1. Methods

Cross-Correlation. Let S_t and W_t be the time series of stock prices or volumes and of web buzz respectively, the cross-correlation between S_t and W_t at the time lag δ is defined as the time lagged Pearson cross correlation coefficient $r(\delta)$:

$$r(\delta) = \frac{\sum_{t=1}^{n}(S_t-\bar{S})(W_{t-\delta}-\bar{W})}{\sqrt{\sum_{t=1}^{n}(S_t-\bar{S})^2}\sqrt{\sum_{t=1}^{n}(W_{t-\delta}-\bar{W})^2}} \quad (1)$$

Where δ goes from -3 to +3, and $\bar{S}$, $\bar{W}$ are the sample averages of the two series[4]. A value δ=0 corresponds to contemporaneous correlation, positive δs correspond to the case in which web buzz tends to anticipate trading variables while the reverse occurs for negative δs. We analyse the cross-correlation for each couple of trading and web variables (overall 96 combinations) and for each type of source and trading market. Tables A.1-A.5 summarize the results.

The **Granger causality test**[5] is used in the literature to assess if a time series X is useful in forecasting another time series Y. If Y is better predicted by the histories of Y and X than with the history of Y only, then X will be said to Granger cause Y. We apply the Granger causality test on each pair of trade and web variables and for each set of sources and stock exchange data available. We test both the null H0: web does not Granger causes stock and the opposite null H0: stock does not Granger causes web and compare the results. Up to three lags for the dependent and the explanatory variables are taken into consideration. As stock returns series could exhibit autoregressive and heteroschedastic error terms, we check the presence of serial correlation and more general ARCH effects on the residuals with the Breusch-Godfrey test and with the ARCH LM test respectively. In all cases considered both tests do not rejects the null (absence of serial correlation and absence of ARCH structure respectively). Tables B1-B5 present the results.

Given that Granger test presupposes stationary series, beforehand we check for unit root with the Augmented Dickey-Fuller test and differentiate series when unit root is not rejected. An additional assumption of Granger causality test is the normality of the error terms. This has been checked with the Jarque-Brera test. Normality of error terms has been found for most of the variables, with occasional exceptions usually related to the stock variable *high(t)-low(t)*.

In order to take into account non linarites and non-Gaussian error terms in the Granger regression, whenever they occur, we perform the **Mann-Whitney U test** (also known as Wilcoxon rank-sum test)[6]. Given two samples of independent observations, this test checks whether the two samples have equal medians. If the null is rejected, one sample will tend to have larger values than the other.

[4] Campbell, Lo, MacKinlay, (1996), *The Econometrics of Financial Markets*. NJ: Princeton University Press.
[5] Granger, C. W. J. (1969), Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica* 37 (3): 424–438.
[6] Mann, H. B. and D., R., Whitney, (1947), On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics* 18 (1): 50–60.

We calculate the U-test for each pair of stock and web variables by bootstrapping the estimated residuals of the two test regressions (trying different bootstrap samples from 10 to 1000 runs each):

$$M_1: stock_t = \alpha + \beta stock_{t-1} + \varepsilon_t \quad (2)$$

$$M_2: stock_t = \alpha + \beta stock_{t-1} + \gamma web_{t-1} + \varepsilon_t \quad (3)$$

$$H_0: \bar{R}^2(M_2) - \bar{R}^2(M_1) = 0 \quad (4)$$

We also test the opposite direction (i.e. we regress web variables onto their lagged values and on lagged values of stocks and calculate $\bar{R}^2$) and compare the results. Detailed results are available in Table B.6.

Principal Component Analysis (PCA)[7] is a non-parametric technique able to regroup variables into factors according to the similarities in their behavior. PCA is usually the first step in the attempt to identify latent (often multidimensional and not directly measurable) constructs in the data. Ideally if all variables behave similarly, PCA would identify a unique single factor capturing a high percentage of the total variance with all variables displaying high loadings (with the same sign). Multiple relevant factors (i.e. with eigenvalue higher that one) are symptom of dissimilar behaviors and would call for further analysis on the drivers of such diversity. Results summarized in Tables C1 and C2.

[7] Jolliffe, I. T. (2002). *Principal Component Analysis,* second edition Springer-Verlag.

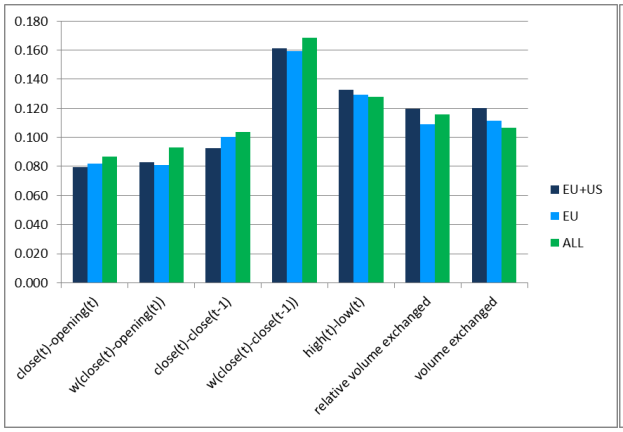
A.2 Selected tables and figures

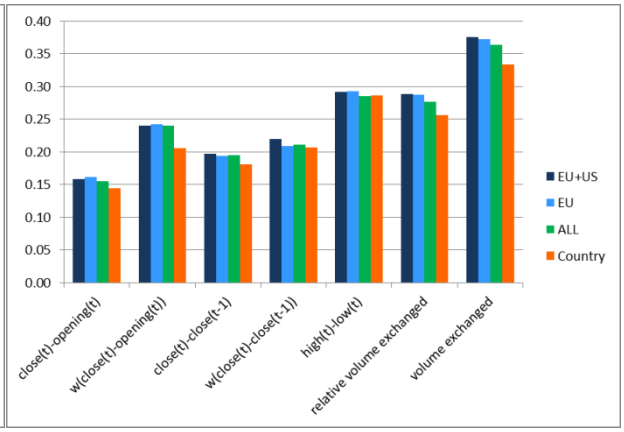
Cross-Correlation: selected tables

Table A. 1. Contemporaneous correlation, average across banks

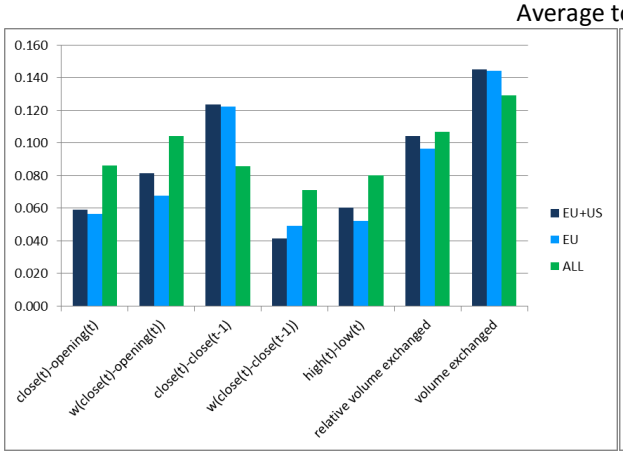
NY stock exchange | Country stock exchange

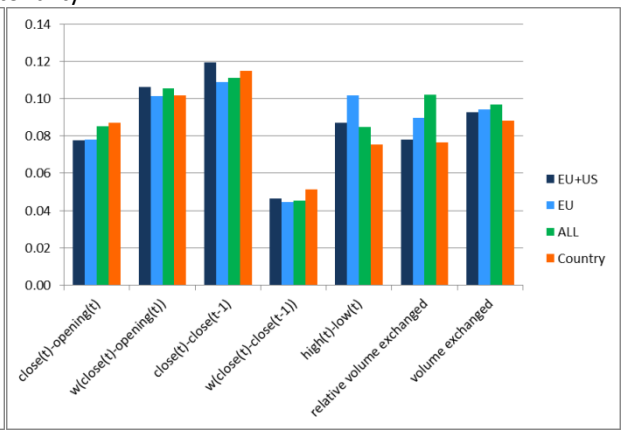
Number of web articles



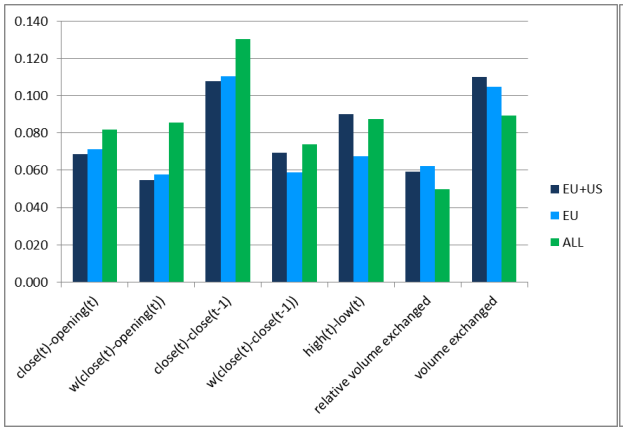


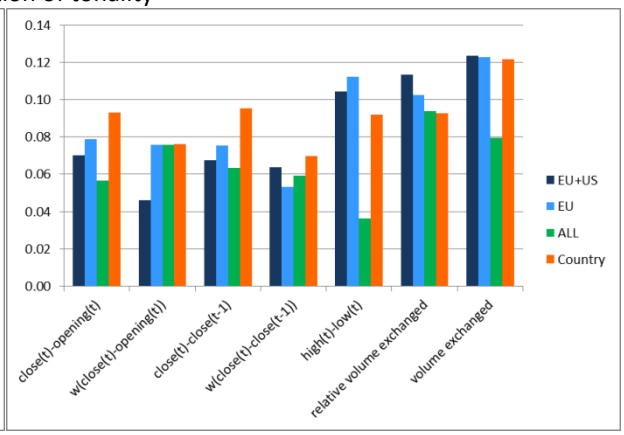
Average tonality



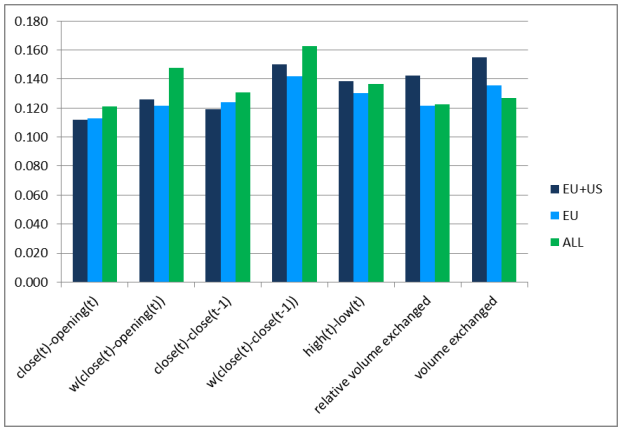


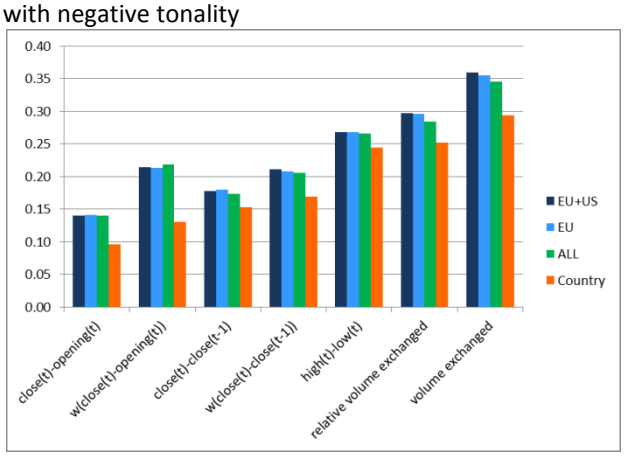
Standard deviation of tonality



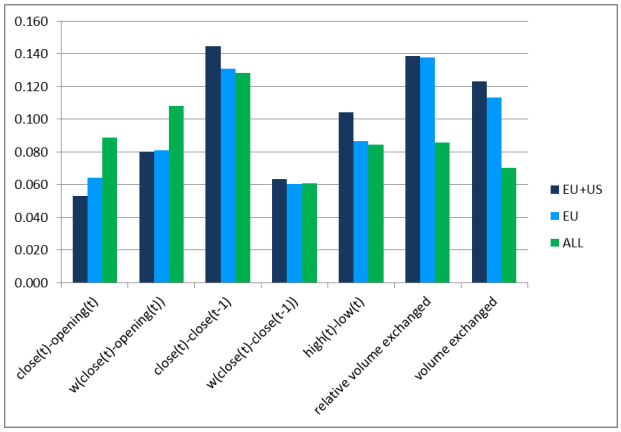


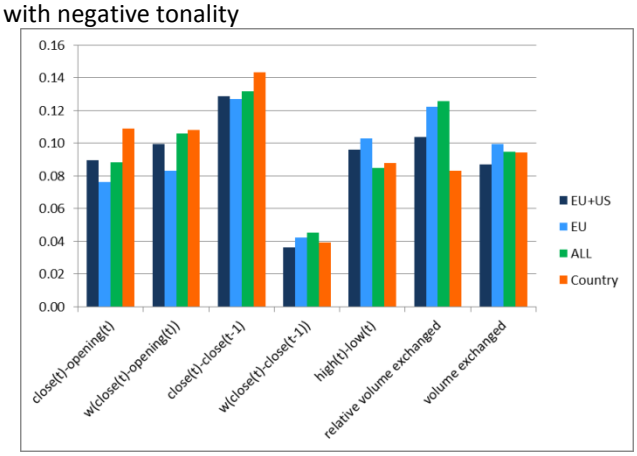
Number of web articles with negative tonality



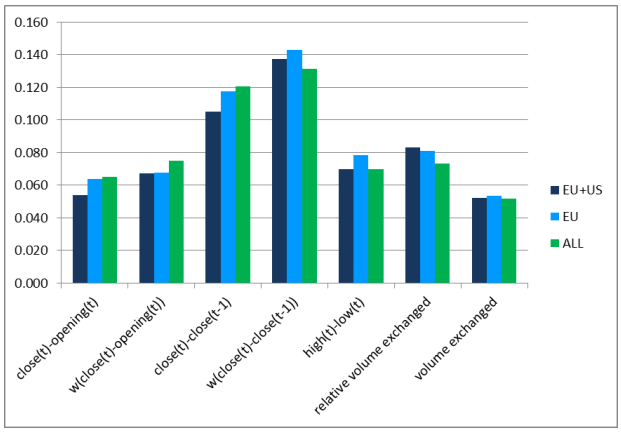


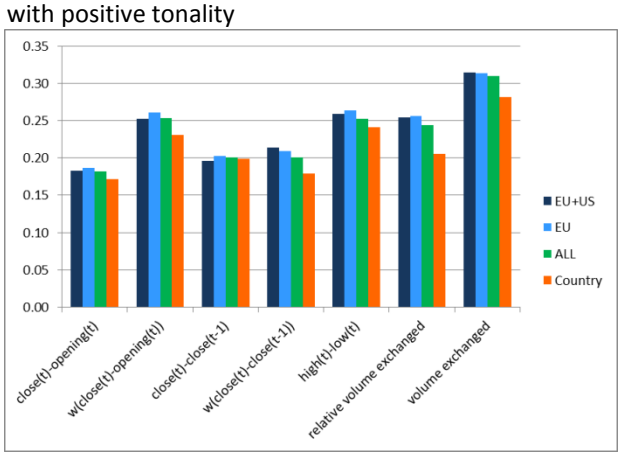
Share of web articles with negative tonality



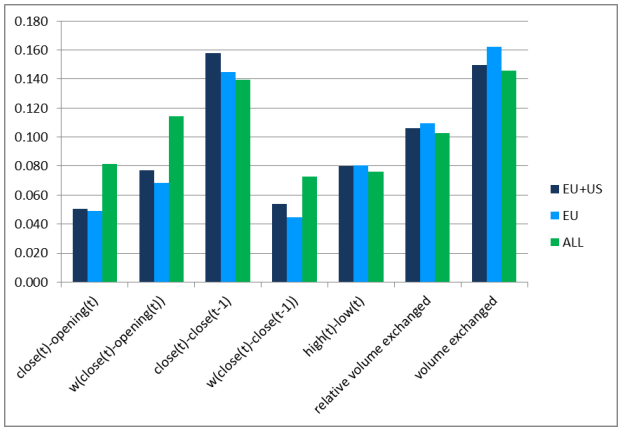


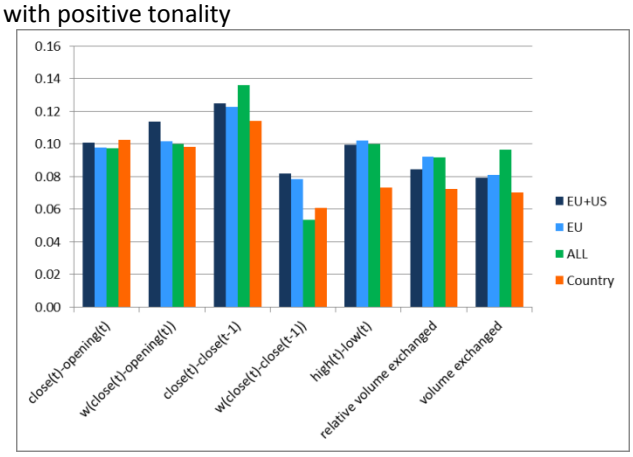
Number of web articles with positive tonality



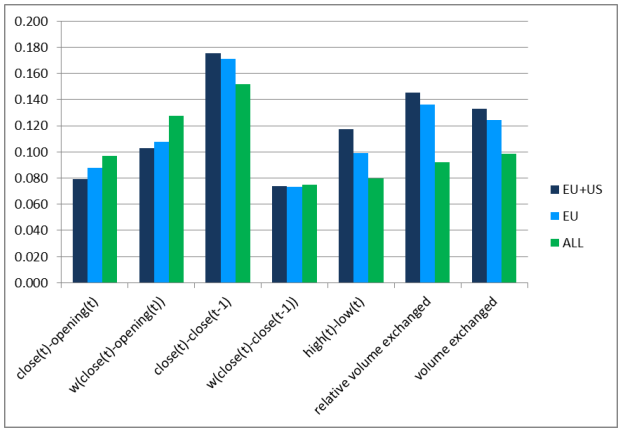


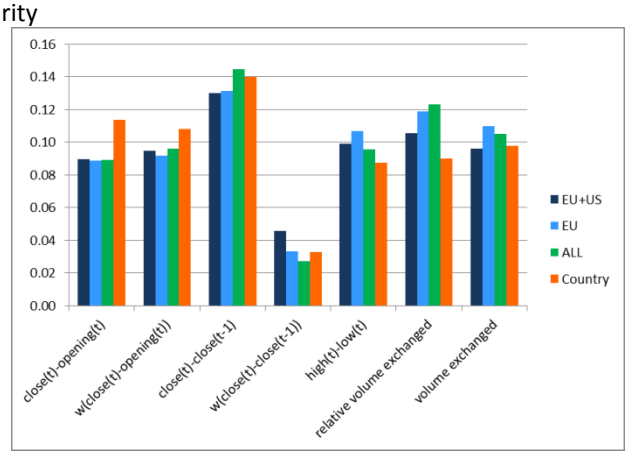
Share of web articles with positive tonality



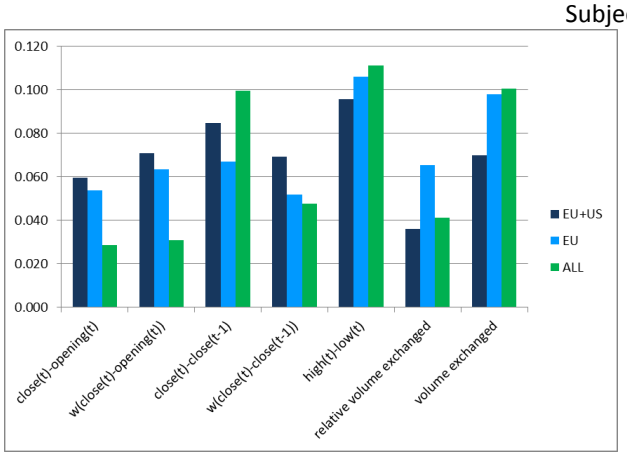


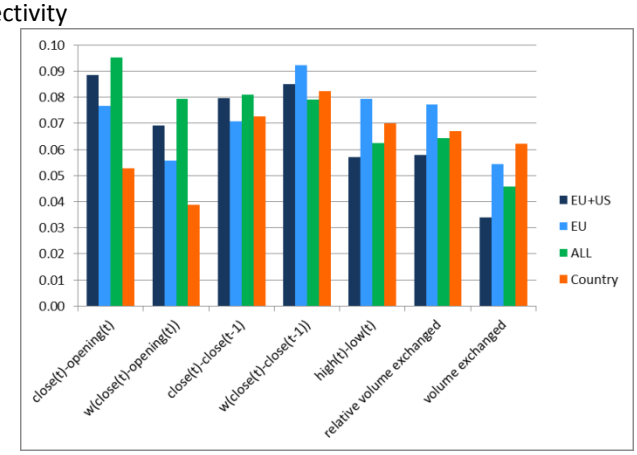
Polarity



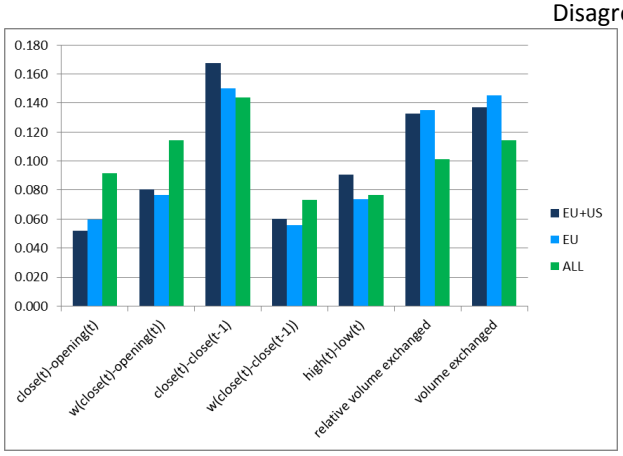


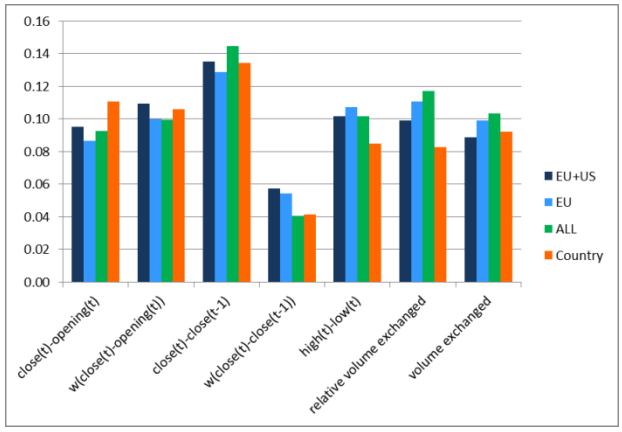
Subjectivity



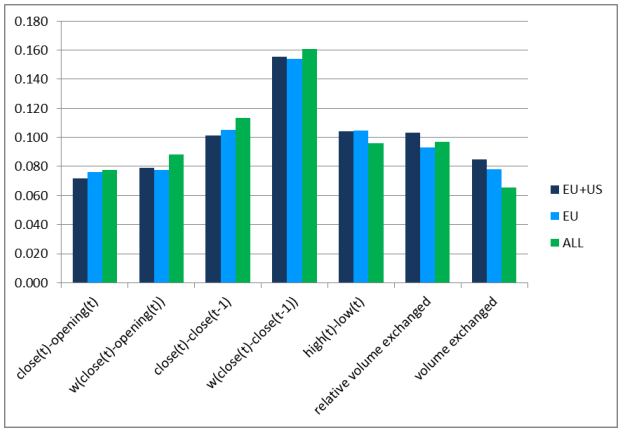


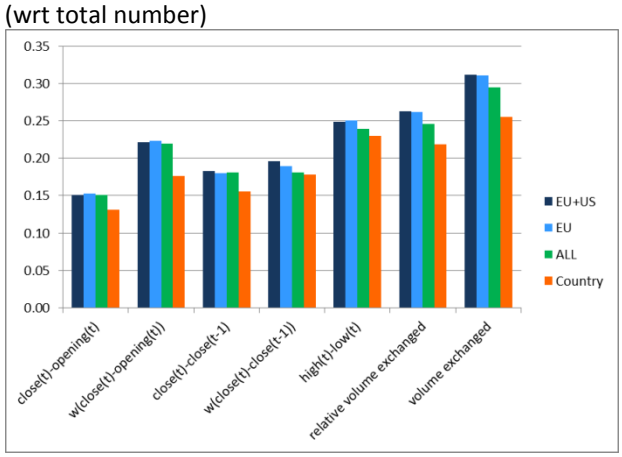
Disagreement



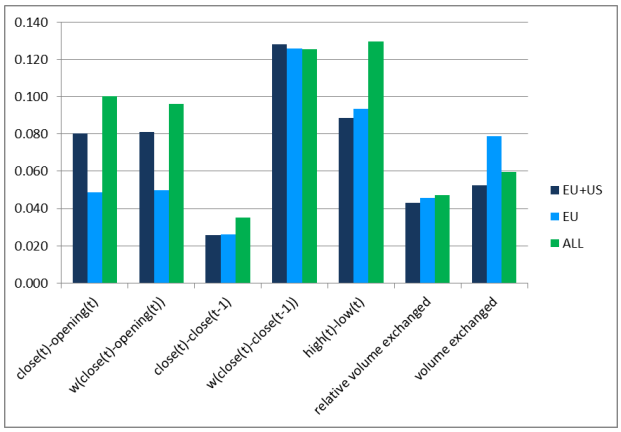


Share of web articles (wrt total number)





Share of web articles (wrt previous day)



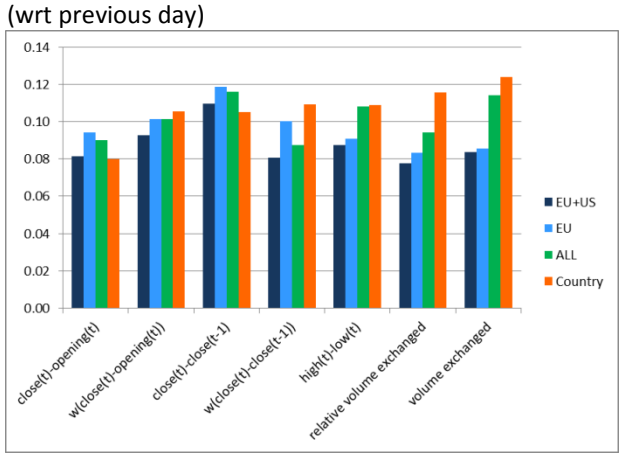


Table A. 2. Cross correlation function between various web variables and measures of stock volume exchanged (data by bank, sources EU+US).

In abscissa the number of lags (0 indicates instantaneous correlation, values with positive sign indicate the cross correlation of stock variables at time t and web variables at time t+lag, negative values indicate the cross correlation of stock variables at time t and web variables at time t-lag).

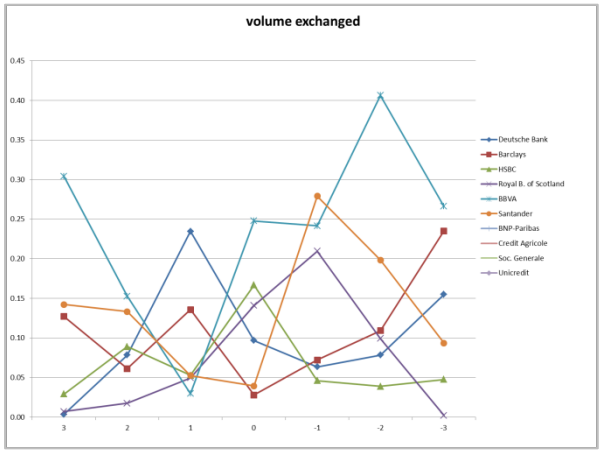
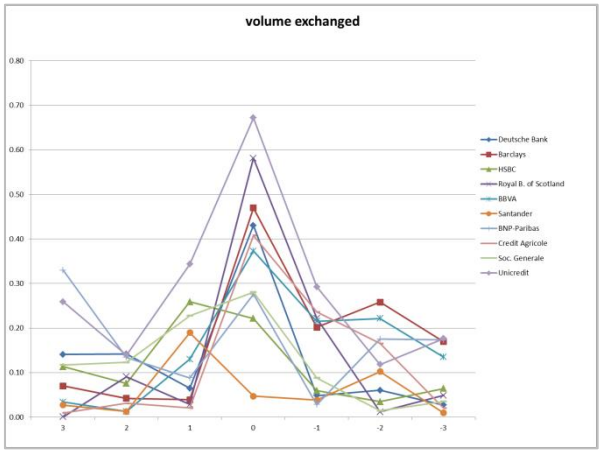
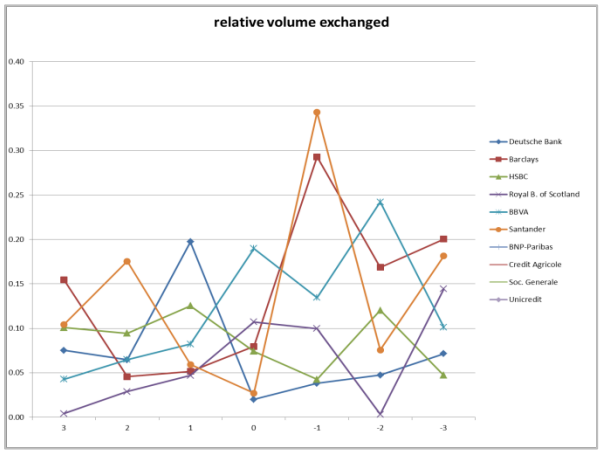
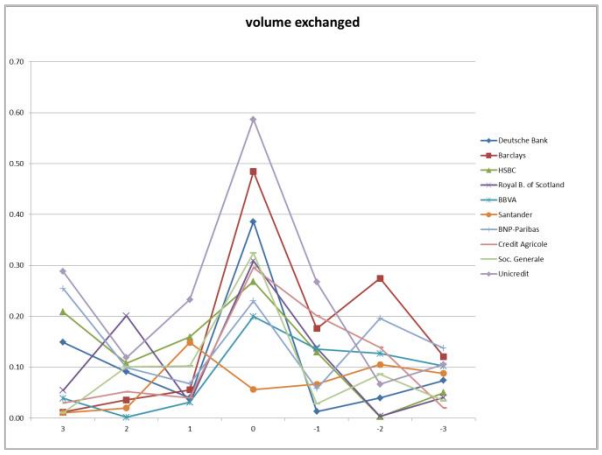
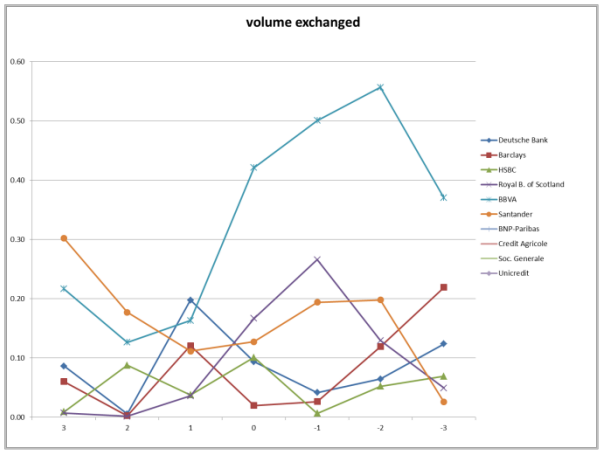
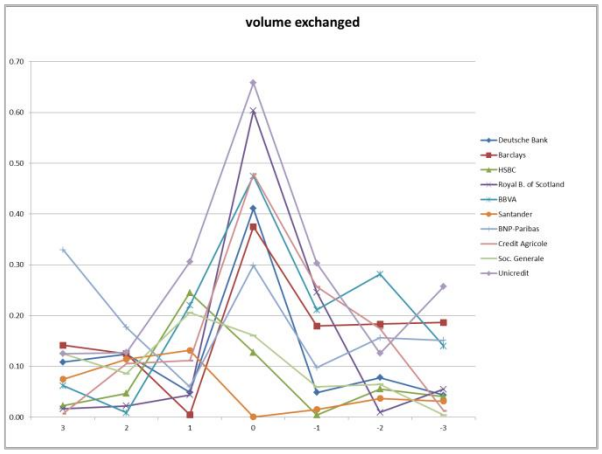
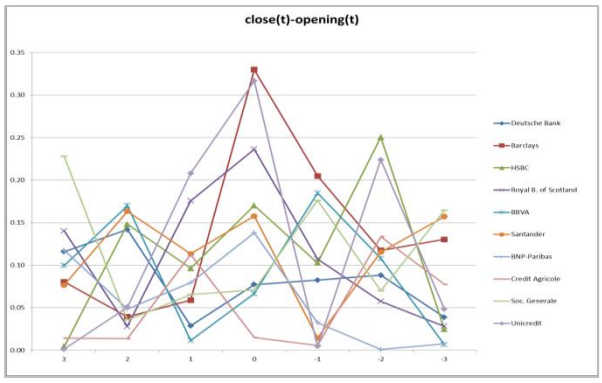
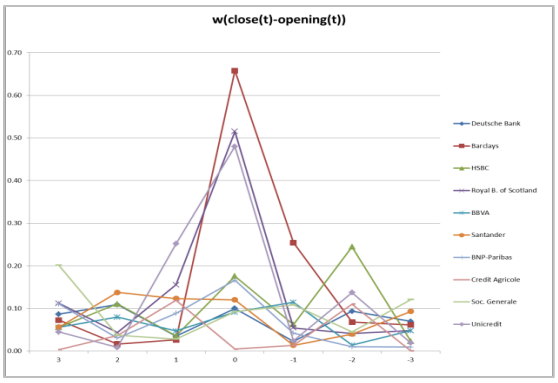
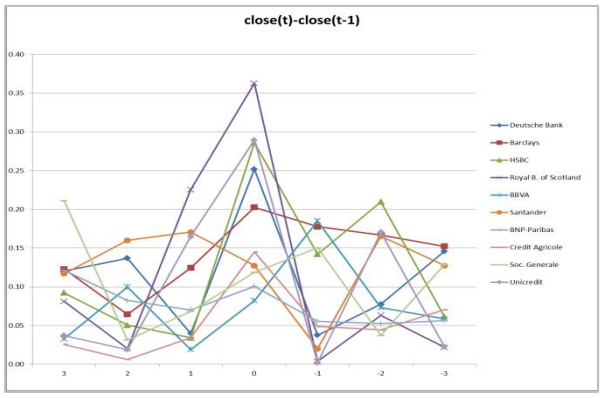
NY stock exchange	Country stock exchange
Number of web articles	
	
Number of web articles with positive tonality	
	
Number of web articles with negative tonality	
	

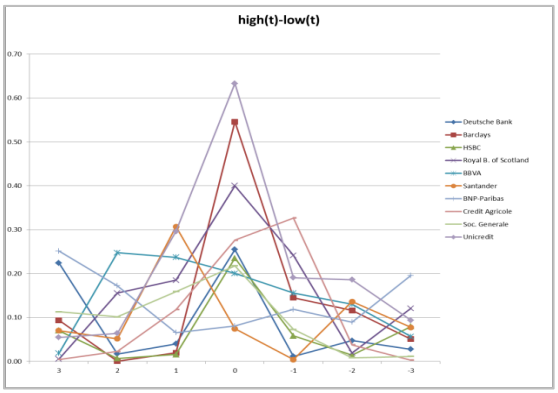
Table A. 3. Cross correlation function between the web variable number of articles and various measures of stock prices and volumes (data by bank, European stock exchanges data, sources EU+US).

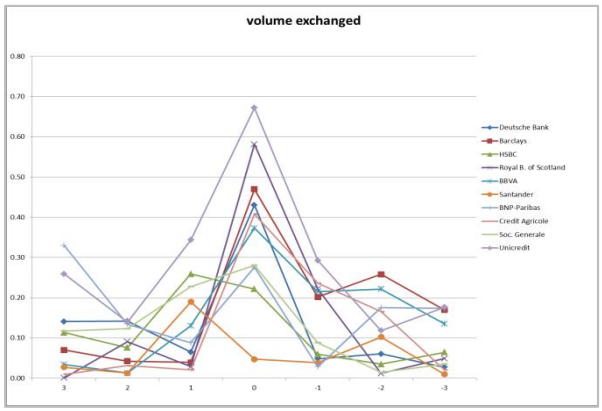
In abscissa the number of lags (0 indicates instantaneous correlation, values with positive sign indicate the cross correlation of stock variables at time t and web variables at time t+lag, negative values indicate the cross correlation of stock variables at time t and web variables at time t-lag).











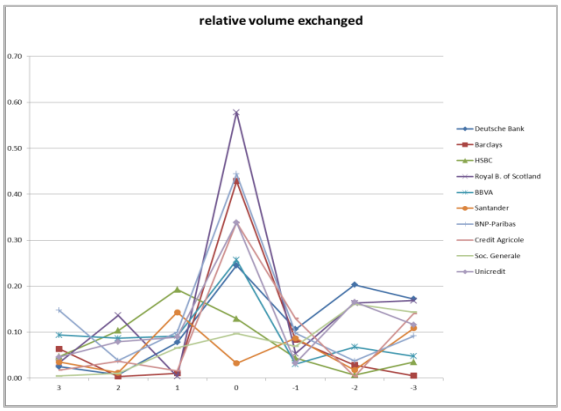


Table A. 4. Cross correlation function between the web variable number of articles and various measures of stock prices and volumes (average across banks, European stock exchanges data, sources EU+US).

Cross correlation function between the web variable *number of articles* and various measures of stock prices and volumes (average across banks, European stock exchanges data, sources EU+US). In abscissa the number of lags (0 indicates instantaneous correlation, values with positive sign indicate the cross correlation of stock variables at time t and web variables at time t+lag, negative values indicate the cross correlation of stock variables at time t and web variables at time t-lag).

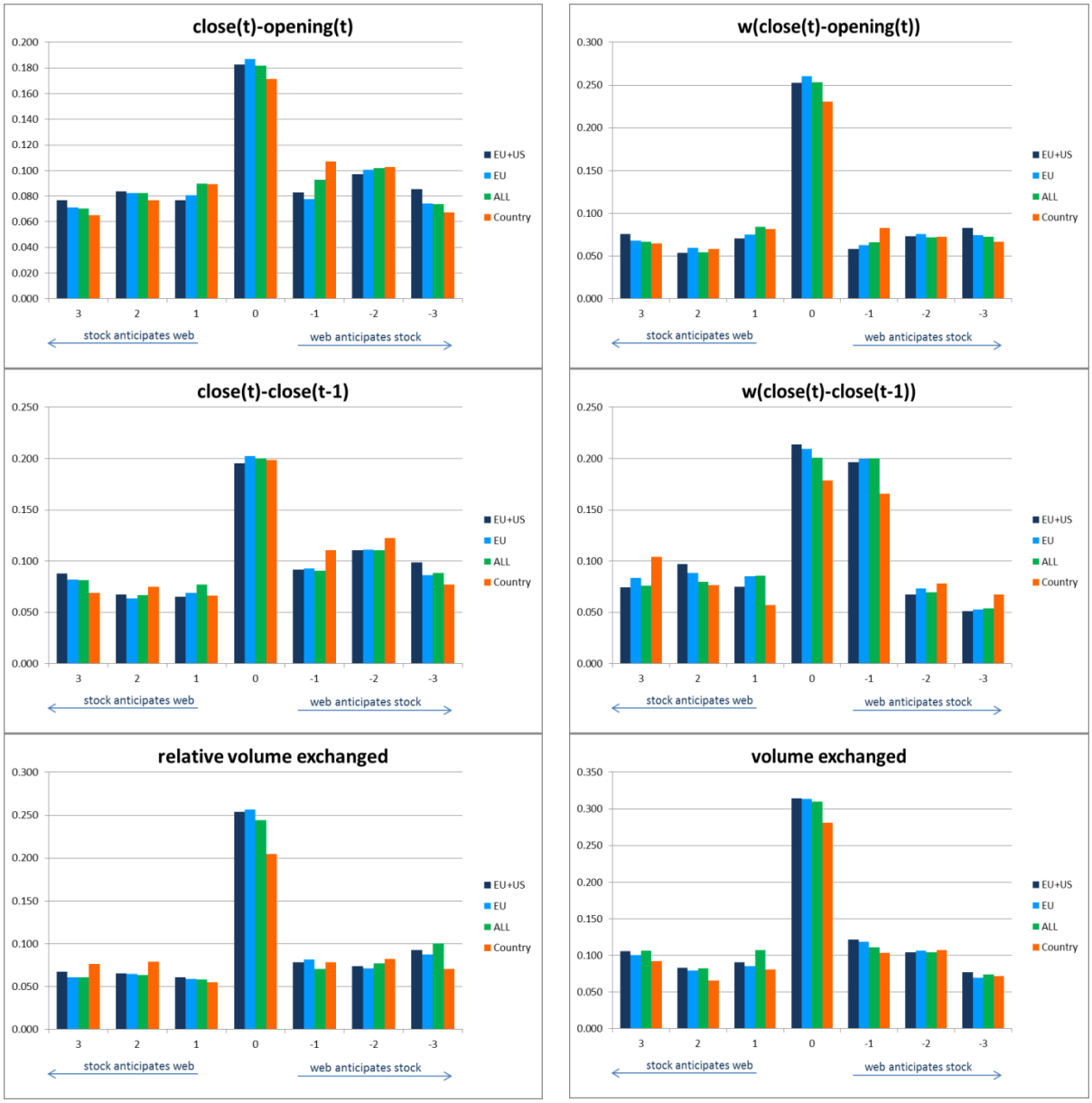
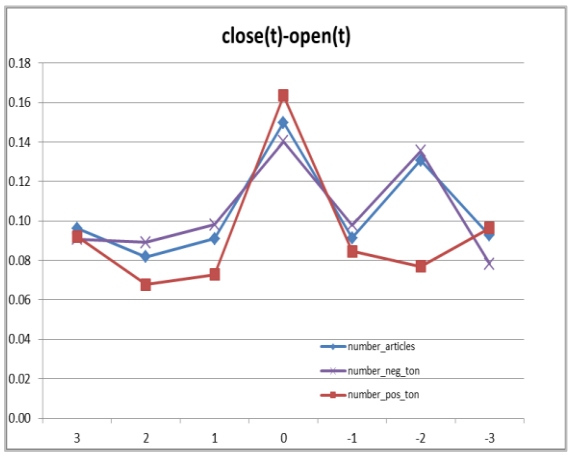


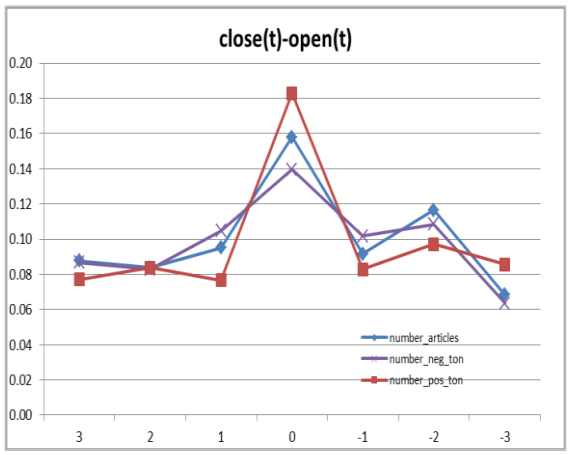
Table A. 5. Cross correlation function between various web variables and measures of stock prices and volume exchanged (average across banks). In abscissa the number of lags (0 indicates instantaneous correlation, values with positive sign indicate the cross correlation of stock variables at time t and web variables at time t+lag, negative values indicate the cross correlation of stock variables at time t and web variables at time t-lag).

Web sources: EU+US-Country

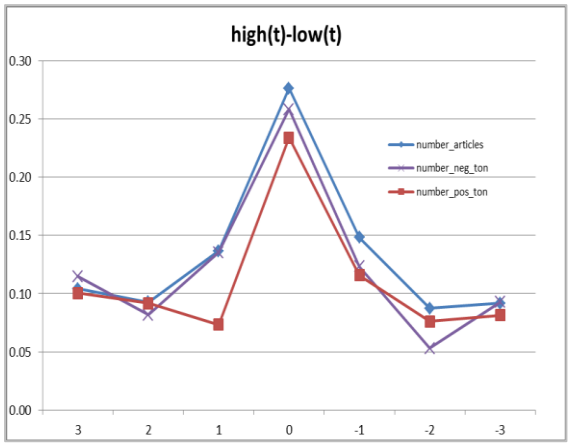
Web sources: EU+US

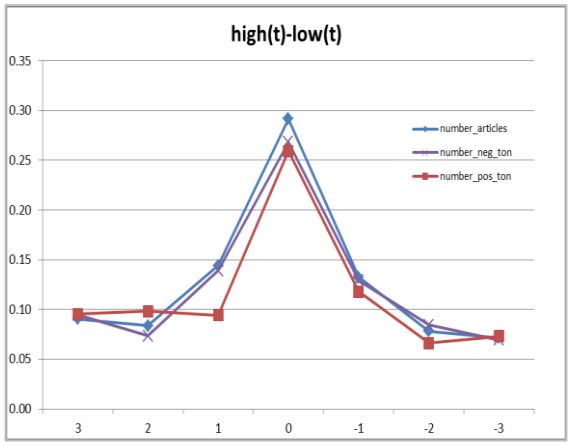
Number of web articles



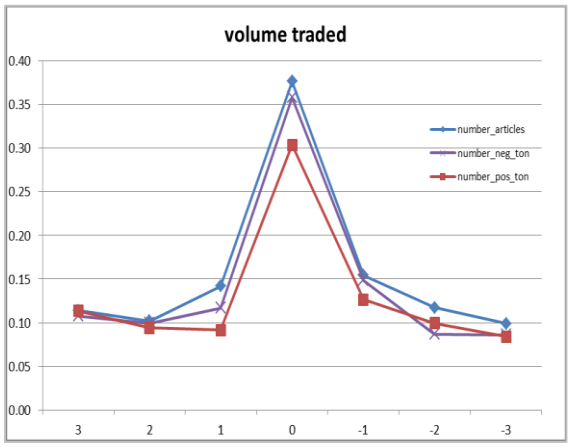


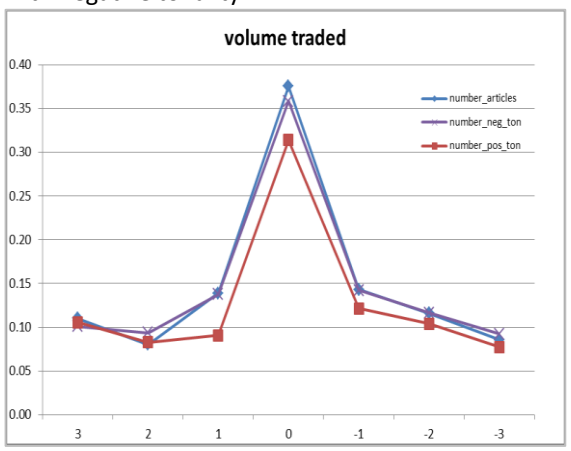
Number of web articles with positive tonality





Number of web articles with negative tonality





Granger causality and U-Rank sum test: selected tables

Table B. 1. Granger test (EU stock exchanges): average results for a selected set of pairs (stock, web). Percentage of banks for which the null hypothesis is rejected at 5% level and percentage reduction in the residual sum of the squares (RSS). Country stock data and various sources; H_0: Web does not Granger cause Stocks (W vs S) or H_0: Stocks does not Granger cause Web (S vs W). Lag of the Granger test = 1 (results with 2 and 3 lags are on average worst so they are not reported). We display in the table only those web variables with the highest average correlation with stock prices and volumes. Granger Causality has been computed for all web variables, all sources and all sock variables and for three lags. The full set of results is available on request.

Country stock exchange data			EU+US sources			ALL sources			EU sources			Country sources		
			number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton
close(t)-opening(t)	W vs.S	% p<5%	0	0	10	0	0	10	0	0	20	10	10	10
		Avg reduction in RSS	NaN	NaN	5%	NaN	NaN	5%	NaN	NaN	5%	5%	6%	5%
	S vs. W	% p<5%	0	20	0	0	10	0	0	20	30	10	0	10
		Avg reduction in RSS	NaN	5%	NaN	NaN	5%	NaN	NaN	5%	5%	5%	NaN	16%
w(close(t)-opening(t))	W vs.S	% p<5%	10	10	0	10	10	0	10	10	10	10	20	0
		Avg reduction in RSS	6%	7%	NaN	5%	6%	NaN	5%	5%	5%	7%	7%	NaN
	S vs. W	% p<5%	10	20	10	10	20	10	10	20	20	10	10	10
		Avg reduction in RSS	9%	4%	6%	8%	5%	4%	8%	5%	6%	5%	9%	16%
close(t)-close(t-1)	W vs.S	% p<5%	0	0	10	10	0	10	0	0	10	0	20	20
		Avg reduction in RSS	NaN	NaN	4%	4%	NaN	4%	NaN	NaN	6%	NaN	4%	6%
	S vs. W	% p<5%	10	10	0	0	10	0	10	20	20	10	10	10
		Avg reduction in RSS	4%	6%	NaN	NaN	5%	NaN	4%	6%	6%	4%	4%	16%
w(close(t)-close(t-1))	W vs.S	% p<5%	30	40	50	30	40	40	30	40	20	30	40	30
		Avg reduction in RSS	11%	10%	6%	12%	10%	8%	12%	11%	6%	7%	6%	7%
	S vs. W	% p<5%	10	10	10	10	10	10	10	0	20	10	10	10
		Avg reduction in RSS	6%	5%	4%	6%	5%	5%	6%	NaN	4%	5%	5%	16%
high(t)-low(t)	W vs.S	% p<5%	10	20	20	10	10	20	10	10	20	10	10	20
		Avg reduction in RSS	9%	5%	6%	7%	7%	5%	9%	6%	5%	9%	7%	5%
	S vs. W	% p<5%	20	40	10	20	40	10	20	30	10	40	20	20
		Avg reduction in RSS	8%	5%	8%	8%	5%	7%	9%	6%	6%	6%	6%	12%
relative vol. exchanged	W vs.S	% p<5%	20	20	30	20	20	20	20	20	20	20	10	20
		Avg reduction in RSS	7%	6%	6%	7%	7%	7%	7%	6%	5%	5%	5%	6%
	S vs. W	% p<5%	10	10	0	10	10	0	0	0	10	30	10	10
		Avg reduction in RSS	4%	4%	NaN	4%	5%	NaN	NaN	NaN	5%	4%	4%	16%
volume exchanged	W vs.S	% p<5%	20	20	20	20	20	20	20	20	10	20	10	20
		Avg reduction in RSS	10%	9%	8%	11%	10%	9%	10%	9%	5%	6%	5%	7%
	S vs. W	% p<5%	10	10	10	10	10	10	10	10	10	20	10	10
		Avg reduction in RSS	7%	6%	4%	8%	6%	6%	7%	5%	5%	5%	4%	16%

Table B. 2. Granger test (EU stock exchanges): results by bank.

Percentage of cases (over all 8 stock variables) in which the null hypothesis is rejected at 5% level and percentage reduction in the residual sum of the squares (RSS). Country stock data, three web variables and various sources; H_0: Web does not Granger cause Stocks (W vs S) or H_0: Stocks does not Granger cause Web (S vs W). Lag of the Granger test = 1.

		EU+US sources			EU sources			ALL sources			Country sources		
Barclays		number_ articles	number_ neg_ton	number_ pos_ton	number_ articles	number_ neg_ton	number_ pos_ton	number articles	number_ neg_ton	number pos_ton	number_ articles	number neg_ton	number_ pos_ton
W vs.S	% p<1%	0	12.5	0	0	0	0	0	0	0	12.5	12.5	0
	% p<5%	12.5	25	12.5	12.5	12.5	12.5	12.5	12.5	12.5	25	50	0
	Avg reduction in RSS	6.3%	5.7%	4.4%	5.5%	5.1%	4.1%	5.5%	5.6%	5.0%	6.1%	5.8%	NaN
S vs. W	% p<1%	12.5	0	0	12.5	0	0	12.5	0	0	0	0	0
	% p<5%	37.5	25	12.5	12.5	0	12.5	12.5	12.5	12.5	25	0	0
	Avg reduction in RSS	6.0%	4.1%	5.7%	8.0%	NaN	5.1%	8.0%	4.3%	4.4%	4.2%	NaN	NaN
BBVA													
W vs.S	% p<1%	0	12.5	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	12.5	37.5	0	12.5	37.5	0	12.5	37.5	0	12.5	37.5
	Avg reduction in RSS	NaN	6.8%	4.6%	NaN	6.2%	4.4%	NaN	4.2%	4.5%	NaN	5.1%	4.6%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	12.5	0
	% p<5%	12.5	25	0	12.5	25	0	12.5	25	0	12.5	50	0
	Avg reduction in RSS	5.1%	4.7%	NaN	5.5%	5.7%	NaN	5.5%	5.0%	NaN	5.4%	5.9%	NaN
BNP Paribas													
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	37.5	0	0	37.5	0	0	12.5	0	0	12.5
	Avg reduction in RSS	NaN	NaN	4.9%	NaN	NaN	5.3%	NaN	NaN	5.3%	NaN	NaN	5.2%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Crédit Agricole													
W vs.S	% p<1%	12.5	0	0	12.5	0	12.5	12.5	12.5	0	12.5	12.5	0
	% p<5%	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5
	Avg reduction in RSS	8.7%	6.4%	6.7%	9.0%	6.4%	7.5%	9.0%	7.1%	5.6%	9.0%	7.5%	4.3%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	12.5	0	0	25	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	4.5%	NaN	NaN	4.0%	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Deutsche Bank													
W vs.S	% p<1%	12.5	0	25	25	25	37.5	25	25	25	0	0	25
	% p<5%	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5
	Avg reduction in RSS	6.4%	6.0%	7.5%	6.9%	7.2%	7.8%	6.9%	6.8%	7.4%	6.3%	5.3%	7.0%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

RSS=Residual Sum of the Squares

Table B.2 follows from previous page

		EU+US sources			EU sources			ALL sources			Country sources		
HSCB		number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0	0	0	12.5
	% p<5%	0	0	0	0	0	0	0	0	0	12.5	0	25
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.7%	NaN	7.1%
S vs. W	% p<1%	12.5	0	0	0	0	0	0	0	0	0	0	100
	% p<5%	37.5	37.5	25	25	12.5	25	25	37.5	25	100	0	100
	Avg reduction in RSS	5.7%	4.9%	4.0%	6.2%	5.0%	4.1%	6.2%	5.1%	5.2%	4.7%	NaN	16.0%
Royal Bank of Scotland													
W vs.S	% p<1%	12.5	12.5	0	12.5	12.5	0	12.5	12.5	0	0	12.5	0
	% p<5%	12.5	12.5	0	12.5	12.5	12.5	12.5	12.5	0	12.5	12.5	0
	Avg reduction in RSS	11.6%	10.4%	NaN	13.1%	15.1%	4.2%	13.1%	11.4%	NaN	5.4%	8.8%	NaN
S vs. W	% p<1%	0	0	0	0	25	0	0	0	0	0	0	0
	% p<5%	0	50	0	25	62.5	0	25	50	0	12.5	25	0
	Avg reduction in RSS	NaN	5.3%	NaN	4.3%	6.2%	NaN	4.3%	4.5%	NaN	4.0%	4.3%	NaN
Santander													
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
S vs. W	% p<1%	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	0	12.5
	% p<5%	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5
	Avg reduction in RSS	11.5%	7.1%	8.1%	11.6%	6.8%	8.2%	11.6%	6.9%	7.2%	10.0%	6.6%	7.1%
Société Générale													
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	12.5	0	0	12.5	0	0	12.5	0	0	12.5
	Avg reduction in RSS	NaN	NaN	5.3%	NaN	NaN	4.8%	NaN	NaN	6.0%	NaN	NaN	5.0%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0	12.5	12.5	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.4%	4.5%	NaN
Unicredito													
W vs.S	% p<1%	37.5	25	25	37.5	25	25	37.5	37.5	37.5	12.5	0	12.5
	% p<5%	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	50	37.5
	Avg reduction in RSS	11.8%	11.0%	8.2%	11.9%	10.4%	8.1%	11.9%	12.0%	9.5%	6.9%	4.6%	6.1%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	12.5	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0%	NaN	NaN	NaN	NaN

RSS=Residual Sum of the Squares

Table B. 3. Granger test (NYSE): average results for a selected set of pairs (stock, web).

Percentage of banks for which the null hypothesis is rejected at 5% level and percentage reduction in the residual sum of the squares (RSS). New York stock data and various sources; H_0: Web does not Granger cause Stocks (W vs S) or H_0: Stocks does not Granger cause Web (S vs W). Lag of the Granger test = 1.

NY stock exchange data			EU+US sources			ALL sources			EU sources		
			number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton	number_articles	number_neg_ton	number_pos_ton
close(t)-opening(t)	W vs.S	% p<5%	16.7	0.0	16.7	16.7	0.0	16.7	16.7	0.0	16.7
		Avg reduction in RSS	5%	NaN	5%	5%	NaN	6%	6%	NaN	4%
	S vs. W	% p<5%	16.7	16.7	16.7	0.0	0.0	16.7	16.7	33.3	33.3
		Avg reduction in RSS	5%	4%	4%	NaN	NaN	4%	7%	6%	5%
w(close(t)-opening(t))	W vs.S	% p<5%	16.7	0.0	16.7	16.7	0.0	16.7	16.7	0.0	16.7
		Avg reduction in RSS	5%	NaN	5%	5%	NaN	5%	5%	NaN	12%
	S vs. W	% p<5%	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0	33.3
		Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	4%	NaN	4%
close(t)-close(t-1)	W vs.S	% p<5%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	S vs. W	% p<5%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7
		Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6%
w(close(t)-close(t-1))	W vs.S	% p<5%	16.7	0.0	16.7	16.7	0.0	16.7	16.7	0.0	16.7
		Avg reduction in RSS	4%	NaN	5%	5%	NaN	6%	4%	NaN	5%
	S vs. W	% p<5%	16.7	16.7	16.7	16.7	16.7	0.0	33.3	16.7	16.7
		Avg reduction in RSS	6%	12%	4%	6%	13%	NaN	6%	15%	6%
high(t)-low(t)	W vs.S	% p<5%	16.7	16.7	16.7	0.0	16.7	33.3	0.0	0.0	16.7
		Avg reduction in RSS	4%	5%	5%	NaN	4%	5%	NaN	NaN	7%
	S vs. W	% p<5%	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	33.3
		Avg reduction in RSS	6%	4%	7%	6%	4%	6%	7%	5%	8%
relative vol. exchanged	W vs.S	% p<5%	0.0	0.0	16.7	0.0	0.0	16.7	0.0	0.0	33.3
		Avg reduction in RSS	NaN	NaN	6%	NaN	NaN	5%	NaN	NaN	4%
	S vs. W	% p<5%	16.7	16.7	16.7	16.7	16.7	0.0	16.7	16.7	33.3
		Avg reduction in RSS	10%	15%	5%	10%	17%	NaN	10%	17%	6%
volume exchanged	W vs.S	% p<5%	0.0	16.7	0.0	16.7	16.7	16.7	0.0	16.7	0.0
		Avg reduction in RSS	NaN	4%	NaN	4%	4%	4%	NaN	5%	NaN
	S vs. W	% p<5%	16.7	16.7	0.0	16.7	16.7	0.0	16.7	16.7	33.3
		Avg reduction in RSS	5%	5%	NaN	5%	6%	NaN	5%	6%	10%

Table B. 4. Granger test (NYSE): results by bank.

Percentage of cases (over all 8 stock variables) in which the null hypothesis is rejected at 5% level and percentage reduction in the residual sum of the squares (RSS). New York stock data, three web variables and various sources; H_0: Web does not Granger cause Stocks (W vs S) or H_0: Stocks does not Granger cause Web (S vs W). Lag of the Granger test = 1.

		EU+US sources			EU sources			ALL sources		
Barclays		number_ articles	number_ neg_ton	number_ pos_ton	number_ articles	number_ neg_ton	number_ pos_ton	number_ articles	number_ neg_ton	number_ pos_ton
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	12.5	0	0	12.5	0	0	12.5
	Avg reduction in RSS	NaN	NaN	5.5%	NaN	NaN	5.5%	NaN	NaN	5.1%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BBVA										
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Deutsche Bank										
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	25	0	37.5	25	0	37.5	25	0	37.5
	Avg reduction in RSS	5.0%	NaN	5.4%	5.3%	NaN	5.9%	5.0%	NaN	5.4%
S vs. W	% p<1%	12.5	25	12.5	37.5	25	12.5	12.5	25	0
	% p<5%	50	50	25	50	62.5	25	50	50	12.5
	Avg reduction in RSS	6.8%	9.1%	5.9%	7.2%	9.6%	5.9%	6.7%	10.2%	5.7%
HSBC										
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	12.5	0	0	12.5	0	0	12.5	12.5
	Avg reduction in RSS	NaN	4.4%	NaN	NaN	5.4%	NaN	NaN	4.0%	4.0%
S vs. W	% p<1%	0	0	0	12.5	12.5	0	0	0	0
	% p<5%	12.5	12.5	12.5	25	12.5	25	0	0	12.5
	Avg reduction in RSS	4.8%	4.0%	4.3%	5.4%	7.4%	4.6%	NaN	NaN	4.2%
Royal Bank of Scotland										
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	12.5	12.5	0	0	0	0	0	12.5	0
	Avg reduction in RSS	4.3%	4.8%	NaN	NaN	NaN	NaN	NaN	4.3%	NaN
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	12.5	12.5	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	4.4%	5.4%	NaN	NaN	NaN	NaN	NaN
Santander										
W vs.S	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	12.5	0	12.5	12.5	0	12.5	25	0	25
	Avg reduction in RSS	4.2%	NaN	5.4%	4.3%	NaN	5.5%	4.7%	NaN	5.1%
S vs. W	% p<1%	0	0	0	0	0	0	0	0	0
	% p<5%	0	0	0	0	0	0	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

RSS=Residual Sum of the Squares

Table B. 5, Granger test: results by bank, sources for web buzz: EU+US-Country.

Percentage of cases (over all 8 stock variables) in which the null hypothesis is rejected at 5% level and percentage reduction in the residual sum of the squares (RSS). Three web variables. H_0: Web does not Granger cause Stocks (W vs S) or H_0: Stocks does not Granger cause Web (S vs W). Lag of the Granger test = 1.

Barclays		number_articles	number_neg_ton	number_pos_ton
W vs.S	% p<1%	0	0	0
	% p<5%	12.5	12.5	12.5
	Avg reduction in RSS	5.4%	5.4%	4.5%
S vs. W	% p<1%	12.5	0	0
	% p<5%	37.5	0	12.5
	Avg reduction in RSS	5.8%	NaN	5.8%
BBVA				
W vs.S	% p<1%	0	0	0
	% p<5%	12.5	12.5	0
	Avg reduction in RSS	4.1%	6.2%	NaN
S vs. W	% p<1%	0	0	12.5
	% p<5%	0	12.5	100
	Avg reduction in RSS	NaN	4.4%	6.2%
BNP Paribas				
W vs.S	% p<1%	0	0	0
	% p<5%	0	25	25
	Avg reduction in RSS	NaN	4.2%	4.7%
S vs. W	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN
Crédit Agricole				
W vs.S	% p<1%	0	0	12.5
	% p<5%	12.5	0	12.5
	Avg reduction in RSS	6.2%	NaN	8.2%
S vs. W	% p<1%	0	12.5	0
	% p<5%	0	100	0
	Avg reduction in RSS	NaN	5.3%	NaN
Deutsche Bank				
W vs.S	% p<1%	0	0	0
	% p<5%	37.5	25	37.5
	Avg reduction in RSS	5.4%	5.5%	6.1%
S vs. W	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN

HSCB		number_articles	number_neg_ton	number_pos_ton
W vs.S	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN
S vs. W	% p<1%	12.5	0	0
	% p<5%	37.5	37.5	0
	Avg reduction in RSS	5.8%	4.7%	NaN
Royal Bank of Scotland				
W vs.S	% p<1%	12.5	12.5	0
	% p<5%	12.5	12.5	0
	Avg reduction in RSS	11.5%	9.3%	NaN
S vs. W	% p<1%	0	0	0
	% p<5%	25	50	0
	Avg reduction in RSS	4.1%	4.8%	NaN
Santander				
W vs.S	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN
S vs. W	% p<1%	37.5	0	12.5
	% p<5%	50	12.5	12.5
	Avg reduction in RSS	7.4%	4.8%	6.8%
Société Générale				
W vs.S	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN
S vs. W	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN
Unicredit				
W vs.S	% p<1%	37.5	37.5	37.5
	% p<5%	37.5	37.5	37.5
	Avg reduction in RSS	13.5%	16.3%	9.2%
S vs. W	% p<1%	0	0	0
	% p<5%	0	0	0
	Avg reduction in RSS	NaN	NaN	NaN

RSS=Residual Sum of the Squares

Table B. 6. Wilcoxon-Mann-Whitney U test: selected results by bank according to the number of bootstraps. Web variables have been extracted from EU+US sources.

Wilcoxon-Mann-Whitney U test			
1000 bootstraps	p-value		p-value
Unicredit (*close(t)-open(t)*)		**BBVA (*high(t)-low(t)*)**	
H0= web does not Granger Cause stock		H0= web does not Granger Cause stock	
number_articles	0.0000	number_articles	0.0000
number_neg_ton	0.0000	H0= stock does not Granger Cause web	
number_pos_ton	0.0000	number_articles	0.1740
H0= stock does not Granger Cause web			
number_articles	0.1093	**Société Générale (*w(close(t)-opening(t))*)**	
number_neg_ton	0.7472	H0= web does not Granger Cause stock	
number_pos_ton	0.7105	number_neg_ton	0.0000
		H0= stock does not Granger Cause web	
Crédit Agricole (*close(t)-open(t)*)		number_neg_ton	0.8657
H0= web does not Granger Cause stock			
share_neg_ton	0.0000	**Deutsche Bank (*high(t)-low(t)*)**	
H0= stock does not Granger Cause web		H0= web does not Granger Cause stock	
share_neg_ton	0.6979	number_pos_ton	0.000
		H0= stock does not Granger Cause web	
		number_pos_ton	0.3997

Wilcoxon-Mann-Whitney U test			
100 bootstraps	p-value		p-value
Unicredit (*close(t)-open(t)*)		**BBVA (*volume exchanged*)**	
H0= web does not Granger Cause stock		H0= web does not Granger Cause stock	
number_articles	0.0000	number_pos_ton	0.0000
number_neg_ton	0.0000	share_pos_ton	0.0000
number_pos_ton	0.0000	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_pos_ton	0.0870
number_articles	0.5867	share_pos_ton	0.9367
number_neg_ton	0.0000		
number_pos_ton	0.0000		
		Royal Bank of Scotland (*close(t)-open(t)*)	
Crédit Agricole (*close(t)-open(t)*)		H0= web does not Granger Cause stock	
H0= web does not Granger Cause stock		number_articles	0.0053
share_neg_ton	0.0000	number_neg_ton	0.0000
H0= stock does not Granger Cause web		H0= stock does not Granger Cause web	
share_neg_ton	0.1430	number_articles	0.1199
		number_neg_ton	0.4763
Deutsche Bank (*high(t)-low(t)*)			
H0= web does not Granger Cause stock		**Société Générale (*w(close(t)-opening(t))*)**	
number_pos_ton	0.000	H0= web does not Granger Cause stock	
H0= stock does not Granger Cause web		number_neg_ton	0.0006
number_pos_ton	0.5405	H0= stock does not Granger Cause web	
		number_neg_ton	0.5357
BBVA (*high(t)-low(t)*)		**Société Générale (*close(t)-close(t-1)*)**	
H0= web does not Granger Cause stock		H0= web does not Granger Cause stock	
number_articles	0.0000	number_articles	0.0000
H0= stock does not Granger Cause web		H0= stock does not Granger Cause web	
number_articles	0.8345	number_articles	0.1597

Wilcoxon-Mann-Whitney U test			
50 bootstraps	p-value		p-value
Unicredit (*close(t)-open(t)*)		**BBVA (*volume exchanged*)**	
H0= web does not Granger Cause stock		H0= web does not Granger Cause stock	
number_articles	0.0000	number_pos_ton	0.0000
number_neg_ton	0.0024	share_pos_ton	0.0000
number_pos_ton	0.4503	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_pos_ton	0.2211
number_articles	0.4340	share_pos_ton	0.0106
number_neg_ton	0.0000		
number_pos_ton	0.0004		
		Royal Bank of Scotland (*close(t)-open(t)*)	
Crédit Agricole (*close(t)-open(t)*)		H0= web does not Granger Cause stock	
H0= web does not Granger Cause stock		number_articles	0.6124
share_neg_ton	0.0312	number_neg_ton	0.0000
share_pos_ton	0.0002	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_articles	0.9533
share_neg_ton	0.3328	number_neg_ton	0.2806
share_pos_ton	0.7433		
		Société Générale (*w(close(t)-open(t))*)	
Deutsche Bank (*high(t)-low(t)*)		H0= web does not Granger Cause stock	
H0= web does not Granger Cause stock		number_neg_ton	0.0096
number_pos_ton	0.0059	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_neg_ton	0.4140
number_pos_ton	0.9368	**Société Générale (*close(t)-open(t)*)**	
		H0= web does not Granger Cause stock	
BBVA (*high(t)-low(t)*)		share_neg_ton	0.0000
H0= web does not Granger Cause stock		H0= stock does not Granger Cause web	
number_articles	0.0001	share_neg_ton	0.1168
H0= stock does not Granger Cause web			
number_articles	0.6716		

Wilcoxon-Mann-Whitney U test			
10 bootstraps	p-value		p-value
Unicredit (w(*close(t)-open(t)*))		**BBVA (*volume exchanged*)**	
H0= web does not Granger Cause stock		H0= web does not Granger Cause stock	
number_articles	0.0890	number_pos_ton	0.0257
number_neg_ton	0.0376	share_pos_ton	0.0452
number_pos_ton	0.3447	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_pos_ton	0.6776
number_articles	0.7337	share_pos_ton	0.9097
number_neg_ton	0.0140		
number_pos_ton	0.0113		
		Royal Bank of Scotland (*close(t)-open(t)*)	
Crédit Agricole (*close(t)-open(t)*)		H0= web does not Granger Cause stock	
H0= web does not Granger Cause stock		number_articles	0.0376
share_neg_ton	0.0113	number_neg_ton	0.0002
share_pos_ton	0.4274	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_articles	0.9097
share_neg_ton	0.5205	number_neg_ton	0.7913
share_pos_ton	0.6776		
		Société Générale (*w(close(t)-open(t))*)	
Deutsche Bank (*high(t)-low(t)*)		H0= web does not Granger Cause stock	
H0= web does not Granger Cause stock		number_neg_ton	0.0376
number_pos_ton	0.3447	H0= stock does not Granger Cause web	
H0= stock does not Granger Cause web		number_neg_ton	0.5708
number_pos_ton	0.8501	**Société Générale (*close(t)-open(t)*)**	
		H0= web does not Granger Cause stock	
BBVA (*high(t)-low(t)*)		share_neg_ton	0.2123
H0= web does not Granger Cause stock		H0= stock does not Granger Cause web	
number_articles	0.9698	share_neg_ton	0.3075
H0= stock does not Granger Cause web			
number_articles	0.9698		

Principal Component: selected tables

Table C. 1. Principal Component Analysis on the entire set of banks: Barclays, BBVA, BNP-Paribas, Crédit Agricole, Deutsche Bank, HSBC, Royal Bank of Scotland, Santander, Société Générale, Unicredit.

PCA on the web variable: number of texts

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	1.56	15.55	15.55	bar_w	-0.432	0.309	-0.442	0.342	0.297	-0.268	-0.132
2	1.33	13.29	28.84	bbva_w	-0.802	-0.025	-0.033	0.100	0.148	0.019	0.098
3	1.28	12.76	41.61	bnp_w	-0.353	-0.367	0.408	-0.387	0.191	0.382	0.173
4	1.07	10.68	52.29	ca_w	0.246	-0.483	-0.533	-0.031	-0.057	0.090	0.061
5	1.00	10.00	62.28	db_w	-0.002	0.500	0.187	0.102	-0.411	-0.642	0.159
6	0.95	9.46	71.74	hsbc_w	0.084	0.051	0.176	0.701	-0.416	0.490	-0.048
7	0.84	8.36	80.11	rbs_w	-0.677	-0.285	-0.142	0.095	-0.344	0.075	0.105
8	0.78	7.79	87.90	san_w	0.034	-0.347	-0.647	-0.187	-0.094	0.236	-0.253
9	0.66	6.57	94.47	sg_w	0.267	-0.379	-0.071	0.441	0.534	0.038	0.482
10	0.55	5.53	100.00	un_w	0.034	0.516	-0.346	-0.223	-0.307	0.075	0.664

PCA on the stock variable: close(t)-open(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	6.04	60.35	60.35	bar_s	-0.605	0.497	0.421	0.313	0.130	0.296	0.036
2	1.00	10.00	70.35	bbva_s	-0.909	-0.135	-0.083	-0.090	0.235	0.021	0.121
3	0.77	7.68	78.03	bnp_s	-0.886	-0.081	-0.072	-0.040	-0.178	-0.109	0.220
4	0.54	5.38	83.42	ca_s	-0.852	-0.187	0.088	0.158	-0.283	0.030	-0.078
5	0.45	4.50	87.92	db_s	-0.795	0.048	0.131	0.194	0.154	-0.442	-0.296
6	0.41	4.14	92.06	hsbc_s	-0.491	0.476	-0.702	0.138	-0.070	0.076	-0.100
7	0.34	3.43	95.49	rbs_s	-0.629	0.472	0.199	-0.559	-0.133	-0.024	-0.104
8	0.23	2.34	97.83	san_s	-0.864	-0.107	-0.126	-0.127	0.394	0.018	0.146
9	0.14	1.40	99.23	sg_s	-0.886	-0.087	0.068	0.111	-0.254	-0.059	0.215
10	0.08	0.77	100.00	un_s	-0.731	-0.472	-0.035	-0.091	-0.019	0.328	-0.312

PCA on the stock variable: high(t)-low(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	4.79	47.94	47.94	bar_s	-0.430	0.101	0.074	0.875	0.104	-0.125	0.083
2	1.06	10.55	58.49	bbva_s	-0.877	-0.038	-0.195	-0.063	-0.045	0.284	0.147
3	0.92	9.17	67.67	bnp_s	-0.847	0.087	-0.141	-0.002	-0.005	-0.072	-0.314
4	0.87	8.70	76.37	ca_s	-0.763	0.117	0.092	-0.262	0.147	-0.357	0.398
5	0.70	6.97	83.34	db_s	-0.776	0.299	-0.156	-0.050	0.124	0.036	-0.199
6	0.55	5.55	88.89	hsbc_s	-0.289	-0.859	-0.268	0.041	-0.103	-0.274	-0.061
7	0.38	3.79	92.68	rbs_s	-0.477	-0.396	0.558	-0.052	0.490	0.238	-0.061
8	0.37	3.74	96.42	san_s	-0.803	-0.121	-0.232	0.070	-0.222	0.370	0.166
9	0.23	2.27	98.69	sg_s	-0.850	0.153	-0.011	-0.139	0.074	-0.232	-0.126
10	0.13	1.31	100.00	un_s	-0.509	-0.003	0.620	-0.011	-0.585	-0.048	-0.056

PCA on the stock variable: volume traded(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	3.83	38.32	38.32	bar_s	-0.345	0.609	0.220	-0.582	-0.010	0.196	0.178
2	1.34	13.39	51.71	bbva_s	-0.787	-0.439	0.020	0.022	-0.046	0.162	0.172
3	1.02	10.16	61.87	bnp_s	-0.718	-0.003	0.031	-0.269	-0.465	-0.119	-0.245
4	0.89	8.93	70.80	ca_s	-0.744	-0.085	0.134	0.335	0.284	-0.035	0.089
5	0.75	7.53	78.33	db_s	-0.619	0.112	-0.223	-0.269	0.481	-0.419	0.139
6	0.65	6.52	84.85	hsbc_s	-0.283	0.385	0.774	0.335	0.016	-0.087	-0.029
7	0.57	5.70	90.55	rbs_s	-0.274	0.577	-0.446	0.400	-0.347	-0.105	0.317
8	0.40	4.03	94.58	san_s	-0.708	-0.449	0.116	-0.057	-0.189	0.118	0.306
9	0.33	3.29	97.87	sg_s	-0.781	0.024	-0.109	0.065	-0.077	-0.245	-0.408
10	0.21	2.13	100.00	un_s	-0.612	0.269	-0.275	0.130	0.243	0.551	-0.246

Table C. 2. Principal Component Analysis on the Euro-area banks : BBVA, Santander, BNP-Paribas, Crédit Agricole, Société Générale, Deutsche Bank, Unicredit.

PCA on the web variable: number of texts

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	1.49	21.32	21.32	bbva_w	-0.760	0.090	0.061	0.340	0.023	0.038	-0.541
2	1.28	18.27	39.60	san_w	-0.679	0.342	0.252	0.211	0.128	-0.124	0.532
3	1.14	16.28	55.87	bnp_w	-0.517	0.093	-0.386	-0.479	-0.528	0.249	0.072
4	0.88	12.60	68.47	ca_w	0.282	0.708	0.105	0.086	0.121	0.621	-0.023
5	0.84	11.99	80.46	sg_w	0.115	0.687	0.315	-0.393	-0.096	-0.462	-0.198
6	0.74	10.63	91.10	db_w	0.273	0.360	-0.541	0.538	-0.397	-0.233	0.047
7	0.62	8.90	100.00	un_w	0.138	-0.210	0.721	0.204	-0.603	0.109	0.024

PCA on the stock variable: close(t)-open(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	5.18	74.06	74.06	bbva_s	-0.920	-0.070	-0.265	-0.079	-0.170	0.054	-0.200
2	0.52	7.38	81.44	bnp_s	-0.892	0.085	0.086	-0.266	0.267	0.217	0.018
3	0.47	6.69	88.13	ca_s	-0.869	0.004	0.350	0.059	-0.317	0.109	0.080
4	0.37	5.25	93.38	db_s	-0.789	0.448	-0.077	0.402	0.096	-0.005	-0.013
5	0.24	3.43	96.81	san_s	-0.872	-0.081	-0.436	-0.081	-0.051	-0.066	0.170
6	0.15	2.09	98.89	sg_s	-0.892	0.130	0.242	-0.219	0.041	-0.280	-0.032
7	0.08	1.11	100.00	un_s	-0.780	-0.530	0.113	0.266	0.161	-0.034	-0.015

PCA on the stock variable: high(t)-low(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	4.39	62.73	62.73	bbva_s	-0.884	0.151	0.281	0.078	0.171	-0.087	-0.273
2	0.81	11.57	74.30	bnp_s	-0.855	0.097	-0.017	-0.077	-0.419	-0.277	0.044
3	0.64	9.13	83.43	ca_s	-0.775	-0.047	-0.448	0.348	0.237	-0.114	0.082
4	0.41	5.87	89.30	db_s	-0.798	0.205	-0.138	-0.485	0.248	0.024	0.073
5	0.39	5.50	94.80	san_s	-0.804	0.086	0.512	0.169	0.032	0.124	0.201
6	0.23	3.25	98.05	sg_s	-0.863	0.019	-0.267	0.041	-0.248	0.337	-0.086
7	0.14	1.95	100.00	un_s	-0.503	-0.852	0.085	-0.113	0.027	-0.015	-0.012

PCA on the stock variable: volume traded(t)

Eigenvalues of correlation matrix				Factor coordinates of the variables, based on correlations							
Factors	Eigenvalue	% Total (variance)	Cumulative (%)	Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	3.64	52.02	52.02	bbva_s	-0.831	0.338	0.225	0.027	-0.091	0.043	0.366
2	0.90	12.86	64.88	bnp_s	-0.709	0.171	-0.531	-0.264	0.018	0.337	-0.051
3	0.70	10.04	74.92	ca_s	-0.747	-0.086	0.356	0.161	0.489	0.172	-0.118
4	0.66	9.47	84.39	db_s	-0.615	-0.429	-0.192	0.581	-0.244	0.063	0.004
5	0.50	7.16	91.55	san_s	-0.744	0.506	0.137	0.072	-0.255	-0.173	-0.269
6	0.36	5.20	96.75	sg_s	-0.781	-0.179	-0.308	-0.124	0.246	-0.429	0.055
7	0.23	3.25	100.00	un_s	-0.591	-0.526	0.306	-0.457	-0.262	0.037	-0.042

European Commission

EUR 27023– Joint Research Centre – Institute for the protection and Security of the Citizen

Title: Does web anticipate stocks? Analysis for a subset of systemically important banks

Author(s): Michela Nardo, Erik van der Goot

Luxembourg: Publications Office of the European Union

2014 – 50 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424

ISBN 978-92-79-44708-2

doi: 10.2788/14130

LB-NA-27023-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

doi: 10.2788/14130

ISBN 978-92-79-44708-2

Publications Office